1-1-2015

The Neural Mechanisms Underlying Visual Target Search

Marino Pagan
University of Pennsylvania, marinopagan@gmail.com

The Neural Mechanisms Underlying Visual Target Search

Abstract
The task of finding specific objects and switching between targets is ubiquitous in everyday life. Searching for a particular object requires our brains to activate and maintain a representation of the target (working memory), identify each encountered object (object recognition), and determine whether the currently viewed object matches the sought target (decision making). The comparison of working memory and visual information is thought to happen via feedback of target information from higher-order brain areas to the ventral visual pathway. However, what is exactly represented by these areas and how do they implement this comparison still remains unknown. To investigate these questions, we employed a combined approach involving electrophysiology experiments and computational modeling. In particular, we recorded neural responses in inferotemporal (IT) and perirhinal (PRH) cortex as monkeys performed a visual target search task, and we adopted population-based read-outs to measure the amount and format of information contained in these neural populations. In Chapter 2 we report that the total amount of target match information was matched in IT and PRH, but this information was contained in a more "explicit" (i.e. linearly separable) format in PRH. These results suggest that PRH implements an "untangling" computation to reformat its inputs from IT. Consistent with this hypothesis, a simple linear-nonlinear model was sufficient to capture the transformation between the two areas. In Chapter 3, we report that the untangling computation in PRH takes time to evolve. While this type of dynamic reformatting is normally attributed to complex recurrent circuits, here we demonstrated that this phenomenon could be accounted by the same instantaneous linear-nonlinear model presented in Chapter 2. This counterintuitive finding was due to the existence of non-stationarities in the IT neural representation. Finally, in Chapter 4 we completely describe a novel set of methods that we developed and applied in Chapters 2 and 3 to quantify the task-specific signals contained in the heterogeneous neural responses in IT and PRH, and to relate these signals to measures of task performance. Together, this body of work revealed a previously unknown untangling computation in PRH during visual search, and demonstrated that a feed-forward linear-nonlinear model is sufficient to describe this computation.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Neuroscience

First Advisor
Nicole C. Rust

Keywords
Neural coding, Neural computation, Neuroscience, Visual target search

THE NEURAL MECHANISMS UNDERLYING VISUAL TARGET SEARCH

Marino Pagan

A DISSERTATION

in

Neuroscience

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2015

Supervisor of Dissertation

_____

Nicole Rust
Assistant Professor of Psychology

Graduate Group Chairperson

_____

Joshua Gold
Professor of Neuroscience

Dissertation Committee:

Yale Cohen, Associate Professor of Otorhinolaryngology

Vijay Balasubramanian, Professor of Physics

Eero Simoncelli, Professor of Neural Science, Mathematics, and Psychology

THE NEURAL MECHANISMS UNDERLYING VISUAL TARGET SEARCH

To Chelsea

ACKNOWLEDGEMENTS

My deepest gratitude goes to my advisor and mentor, Nicole Rust. Her passion and dedication have been a constant source of inspiration, and I am forever indebted to her for the incredible amount of support, encouragement and guidance she provided since the beginning of my scientific career.

I would also like to thank the members of my Dissertation Committee, Josh Gold, Yale Cohen, Vijay Balasubramanian and Eero Simoncelli, for all the invaluable advice and support.

In addition, I am very grateful to all the amazing people I had the pleasure to work with in the Rust lab: Jennie Deutsch, Margot Wohl, Krystal Henderson, Noam Roth and Alexandra Smolyanskaya. Outside the Rust lab, the whole Neuroscience community at Penn has also been an incredible source of support and insight.

Finally, but most importantly, I am deeply thankful to my wife Chelsea for all her love, support and patience.

ABSTRACT

THE NEURAL MECHANISMS UNDERLYING VISUAL TARGET SEARCH

Marino Pagan

Nicole Rust

The task of finding specific objects and switching between targets is ubiquitous in everyday life. Searching for a particular object requires our brains to activate and maintain a representation of the target (working memory), identify each encountered object (object recognition), and determine whether the currently viewed object matches the sought target (decision making). The comparison of working memory and visual information is thought to happen via feedback of target information from higher-order brain areas to the ventral visual pathway. However, what is exactly represented by these areas and how do they implement this comparison still remains unknown. To investigate these questions, we employed a combined approach involving electrophysiology experiments and computational modeling. In particular, we recorded neural responses in inferotemporal (IT) and perirhinal (PRH) cortex as monkeys performed a visual target search task, and we adopted population-based read-outs to measure the amount and format of information contained in these neural populations. In Chapter 2 we report that the total amount of target match information was matched in IT and PRH, but this information was contained in a more “explicit” (i.e. linearly separable) format in PRH. These results suggest that PRH implements an “untangling” computation to reformat its inputs from IT. Consistent with this hypothesis, a simple linear-nonlinear model was sufficient to capture the transformation between the two areas. In Chapter 3, we report that the untangling computation in PRH takes time to evolve. While this type of dynamic

reformatting is normally attributed to complex recurrent circuits, here we demonstrated that this phenomenon could be accounted by the same instantaneous linear-nonlinear model presented in Chapter 2. This counterintuitive finding was due to the existence of non-stationarities in the IT neural representation. Finally, in Chapter 4 we completely describe a novel set of methods that we developed and applied in Chapters 2 and 3 to quantify the task-specific signals contained in the heterogeneous neural responses in IT and PRH, and to relate these signals to measures of task performance. Together, this body of work revealed a previously unknown untangling computation in PRH during visual search, and demonstrated that a feed-forward linear-nonlinear model is sufficient to describe this computation.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER 1: Introduction

The problem of visual target search

The ability to search for specific objects is fundamental in everyday life. Throughout our day, we often need to look for a particular item, such as our car keys. Upon having found them, we can rapidly switch to a different target, such as our wallet. Humans are capable to easily solve this task with remarkable speed and accuracy. However, clinical conditions can severely deteriorate our ability to remember and locate familiar objects, as in the case of individuals affected by dementia (Sahgal, Galloway et al. 1992). The goal of the research presented in this dissertation was to investigate the neural underpinnings of our ability to find specific objects and flexibly switch between different targets.

To inform the search for the neural mechanisms underlying flexible target search, it is useful to conceptualize this problem by subdividing it into three elementary components. First, a remembered representation of the target must be activated and maintained in working memory. Second, it is necessary to process visual information to determine the identity of each encountered object. Third, visual and working memory representations must be combined to determine whether a currently viewed object matches the sought target. The next two sections review the current literature on the

brain mechanisms responsible for processing visual information and combining it with target information during object search.

Object recognition mechanisms in the ventral visual pathway

A large number of studies have shed light on the crucial role played by the ventral visual pathway in identifying currently viewed objects (DiCarlo, Zoccolan et al. 2012). As shown in Figure 1a, the ventral visual pathway is composed by a hierarchy of cortical areas in the occipital and temporal lobe, including primary visual cortex (V1), secondary visual cortex (V2), area V4, and culminating in a group of areas called inferotemporal cortex (IT) in the monkey (Felleman and Van Essen 1991), and lateral occipital complex (LOC) in the human (Grill-Spector, Kushnir et al. 1998).

Responses of neurons along the pathway display an increasing selectivity for complex shapes and an increasing invariance to small changes in their position, size and clutter (Kobatake and Tanaka 1994, Ito, Tamura et al. 1995, Hung, Kreiman et al. 2005). Also, receptive fields become larger as visual information is pooled across wider portions of the visual field (Kobatake and Tanaka 1994). In parallel with the incremental complexity of single neuron responses, recent studies have demonstrated that the combined population responses at each successive stage also carry a progressively refined encoding of visual information (Rust and DiCarlo 2010), culminating in a robust representation of the identity of currently viewed objects in IT (Hung, Kreiman et al. 2005).

The increasingly complexity of neural responses along the ventral visual pathway is mirrored by the types of visual deficits produced by lesions or inactivation of its different stages. For example, lesions in V1 cause complete blindness in a well-defined portion of the visual field (Stoerig and Cowey 1997), lesions in V2 and V4 prevent the detection of conjunctions of simple features (Merigan, Nealey et al. 1993, Schiller 1995), while lesions of IT impair the ability to recognize complex objects, such as faces (Yaginuma, Niihara et al. 1982, Holmes and Gross 1984, Schiller 1995, Horel 1996).

Despite not being purely visual, several studies have suggested that an additional area, perirhinal cortex (PRH) should also be considered as part of the ventral visual pathway, because of its role in object visual processing (Murray and Bussey 1999, Bussey and Saksida 2005, Buckley and Gaffan 2006, Baxter 2009, Suzuki and Baxter 2009). PRH is located next to IT in the temporal lobe, and it receives most of its input from IT (Suzuki 1996). As a consequence, PRH neurons also respond strongly to visual stimuli (Nakamura, Matsumoto et al. 1994, Naya, Yoshida et al. 2003, Lehky and Tanaka 2007).

Processing of visual and cognitive signals during target search

In contrast to the extensive literature on processing and representation of visual information, much less is known about how such information is employed to perform actual tasks, such as searching for specific objects. Two classes of tasks have traditionally been used to study visual search. In the first class, subjects are required to

find a particular target within a static scene that was degraded by clutter or lack of contrast, as in the “Where’s Waldo?” puzzles (Desimone and Duncan 1995, Awh, Armstrong et al. 2006). In the second class, subjects are presented with a sequential stream of distinct images, and they are instructed to respond when the currently presented image corresponds to the sought target. In this dissertation we focus on the second class of tasks, and in particular on the most commonly used one, known as Delayed-Match-to-Sample (DMS) task (Mishkin, Prockop et al. 1962, Gaffan 1974, Mishkin and Delacour 1975). In this task, which is designed to model the sequential search commonly occurring when looking for an object, subjects are first cued with a sample image of the sought target. The target then disappears and, after a temporal delay, subjects are required to respond to images that match the target, but not to distractor images.

Because of the temporal delay between the presentation of a sample of the target and the subsequent presentation of match and target stimuli, solving DMS requires to actively hold the target in working memory. Although the exact brain structures and mechanisms underlying working memory are still the subject of active debate (Curtis and Lee 2010, Barak and Tsodyks 2014), the brain area most often implicated is prefrontal cortex (PFC) (Barak, Tsodyks et al. 2010). The major supporting evidence for the role of PFC in maintaining working memory is the experimental finding that neurons in PFC exhibit sustained responses that are selective for different targets even after the disappearance of the target, a phenomenon known as persistent activity (Romo, Brody et al. 1999). Interestingly, this persistent signal shows remarkable dynamics, with different groups of neurons carrying the information at different times (Brody, Hernandez et al. 2003). The neural mechanism generally proposed to underlie

persistent activity features multiple groups of neurons characterized by recurrent excitation and mutual inhibition (Machens, Romo et al. 2005). After the initial activation of the group associated with the active target, recurrent excitatory synapses act to maintain the sustained activity, while inhibitory connections prevent other groups from becoming active as well. Despite the general prevalence of the theory of persistent activity through recurrent connections, it is worth noting the existence of alternative hypotheses postulating the involvement of short-term synaptic plasticity in the maintenance of working memory (Mongillo, Barak et al. 2008, Sugase-Miyamoto, Liu et al. 2008).

Where in the brain do target-specific signals combine with visual information? Numerous sources of evidence suggest that working memory information is fed-back directly into the same areas in the ventral visual pathway that are involved in visual processing, and in particular V4, IT and PRH (Figure 1b). First of all, these areas receive strong inputs from PFC (Markov, Ercsey-Ravasz et al. 2012). The functional role of these projections was first directly demonstrated using monkeys who underwent the resection of posterior corpus callosum and anterior commissure, leaving intact only the anterior corpus callosum, which connects the prefrontal cortices (Tomita, Ohbayashi et al. 1999). In these animals, neurons in IT and PRH were shown to respond to ipsilateral visual cues, a fact that can only be explained by top-down signals from PFC, since visual information could only cross the hemispheres through the anterior corpus callosum.

Further evidence of the importance of IT and PRH for target search is provided by lesion studies. Selective lesions of these brain areas produced significant deficits during DMS tasks (Gaffan and Murray 1992) as well as delayed-nonmatch-to-sample

(DNMS) tasks (Buffalo, Ramus et al. 1999, Buffalo, Ramus et al. 2000). Lesions of PRH produced marked deficits when monkeys had to remember targets over long delays (~1 min), while lesions of IT impaired performances even in the case of very short delays (~0.5 seconds). However, the tasks used in these studies did not required the monkeys to actively maintain a specific target in their working memory, and could be solved simply by looking for repeated stimuli, a strategy that monkeys are known to naturally adopt (Miller and Desimone 1994).

In parallel with the evidence from anatomical and lesion studies, neural recordings of V4, IT and PRH during DMS tasks demonstrated that neurons in these areas are not only modulated by what monkeys are looking at, but also by what monkeys are looking for, a phenomenon known as “feature-based attention” (Maunsell and Treue 2006). Recordings in V4 during DMS revealed that 40% of the neurons were significantly modulated by the identity of a target (Haenny, Maunsell et al. 1988, Maunsell, Sclar et al. 1991). The most prevalent effect among these neurons was an overall increase of responses to “target match” conditions (i.e. conditions in which the viewed stimulus matched the sought target) as compared to “non-match” conditions (i.e. where the identity of target and viewed stimulus differed). Analogously to V4, strong target modulations during DMS were reported also in IT and PRH (Eskandar, Richmond et al. 1992, Miller, Li et al. 1993, Miller and Desimone 1994, Miller, Erickson et al. 1996). In these areas, however, two subpopulations of neurons with opposite response properties were described: one with enhanced responses to target matches, and the other with suppressed responses to target matches.

Despite the great impact of these early studies on our understanding of the brain mechanisms underlying target search, many questions still remain open. First of all, what are the rules of combination of visual and working memory signals? We now know that these computations are implemented in the highest stages of the ventral visual pathway, but their specific implementation is still largely unknown. Second, these areas are known to encode an explicit representation of currently viewed objects: how is this visual representation modulated by cognitive signals? Third, while most studies focused on the modulation of responses to target matches, neural populations in V4, IT and PRH were generally described as largely heterogeneous, and they included cells responding only to specific non-match conditions (Haenny, Maunsell et al. 1988, Eskandar, Richmond et al. 1992): what is the role and significance of these neurons, and how does heterogeneity impact on the population representations in these areas? Fourth, experiments conducted in IT and PRH generally did not differentiate between these two areas, thus lumping together separate neural populations that could potentially express significant differences in their responses. Finally, while it is now well established that combinations of visual and target information exist in V4, IT and PRH, very little is known about the direction of the flow of information, and about the specific computations performed by each area.

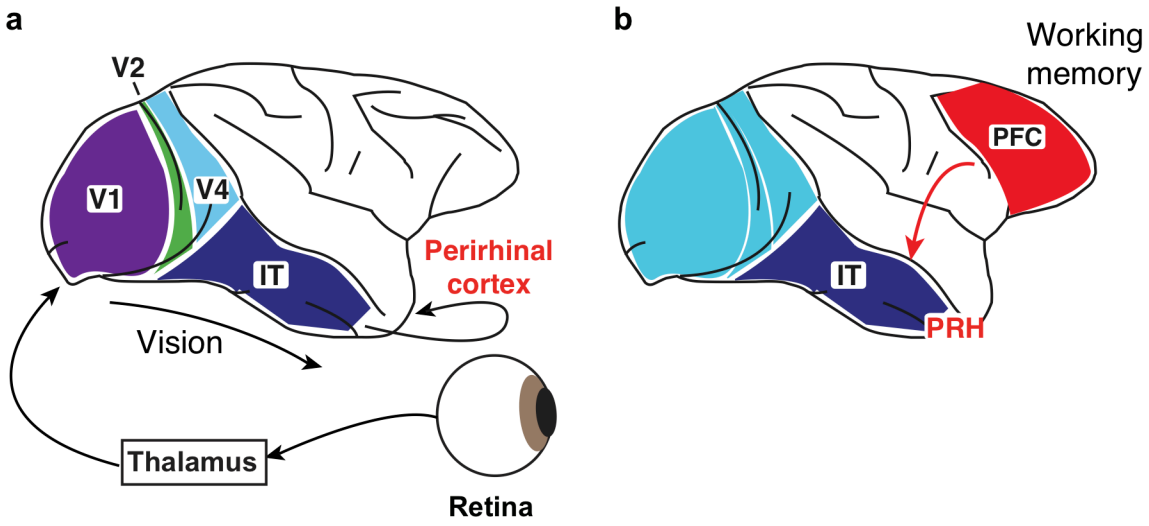

Figure 1-1. *Proposed pathways for processing of visual and working memory information during visual target search.* **a)** Visual information is first encoded in the retina, from where it reaches the primary visual cortex through the lateral geniculate nucleus of the thalamus. Information about the identity of viewed objects is then extracted along the ventral visual pathway (or "what" pathway), composed by V2, V4 and IT. Adjacent to IT, perirhinal cortex constitutes an extension of the classical ventral visual pathway. **b)** Working memory information is generally thought to be maintained in prefrontal cortex. Multiple sources of evidence suggest a top-down projection of this signal to the highest stages of the ventral visual pathway (V4, IT and perirhinal) during target search (Tomita, Ohbayashi et al. 1999, Markov, Ercsey-Ravasz et al. 2012).

Computational aspects of visual target search and the "untangling" framework

Several theoretical models have been proposed for how the brain might implement visual target search (Salinas 2004, Sugase-Miyamoto, Liu et al. 2008, Salinas and Bentley 2009, Engel and Wang 2011). Despite differences in their specific

implementation, all these models share the same general multi-stage structure. In the first stage of this network, visual and working memory signals are combined nonlinearly (e.g. multiplicatively) to produce a target-modulated visual representation. Notably, most existing models make the simplifying assumption that visual and target inputs are "aligned" when first combined, even though neural recordings suggest the presence of misaligned combinations (Haenny, Maunsell et al. 1988, Eskandar, Richmond et al. 1992). This neural population expressing combinations of visual and target information is then thought to send its projections to a higher "decision" brain area, which will eventually produce the solution of the task. Such final output can be represented by a hypothetical "target-present" neuron, whose response is enhanced by any target match condition and suppressed by any non-match condition.

From a computational standpoint, creating such output neuron is equivalent to performing the inverse of an "exclusive OR" (XOR) operation, a fundamental and well-studied concept in computer science and electrical engineering (Horowitz, Hill et al. 1989). In the simple case with two binary inputs, an inverse XOR rule prescribes that the output should be active if an only if the inputs are matched (i.e. both equal to 1, or both equal to 0), and the output should be inactive if the inputs are mismatched (i.e. one is 1 and the other is 0), thus modeling the same match/non-match behavior of the hypothetical "target-present" neuron. Importantly, XOR constitutes the simplest computation that cannot be performed by a classic single-layer network (Minsky and Papert 1969), thus requiring at least two layers of computations to be solved.

A useful way to conceptualize computations requiring multi-stage networks is constituted by the "untangling" framework, which was initially proposed in the context of

object recognition (DiCarlo and Cox 2007). In the case of object recognition, all information necessary to discriminate different objects is already present at the level of the retina. However, the “read-out” necessary to extract this information is extremely complex and nonlinear, causing this information to be effectively inaccessible to downstream neurons. What makes retinal information hard to decode is the fact that representations of different objects are tightly “tangled” in the space of retinal population responses, thus requiring highly nonlinear decision boundaries to discriminate between different objects. From this perspective, the goal of the ventral visual pathway is to progressively reformat the retinal representation into an explicit, or “untangled” format that can be easily read-out by downstream neurons. The same idea can be adopted to describe the computations required during target search, where initial mixtures of visual and target signals represent the “target-present” signal in a tangled manner, thus requiring at least one reformatting computation.

Besides object recognition and target search, the untangling framework can be extended to a large class of perceptual and cognitive tasks. For example, discrimination of complex auditory stimuli also requires a series of computational stages to reformat the initial cochlear representation (Sharpee, Atencio et al. 2011), and similar considerations have been equally made in relation to the processing of olfactory stimuli (Rokni, Hemmelder et al. 2014). In addition to perceptual computations, many cognitive tasks can be interpreted in terms of changes of representation. An example is provided by language processing, where the initial acoustic representation of words must be transformed into a semantic representation to formulate appropriate responses. Recognizing the fact that all these computations share the same untangling framework is advantageous for two reasons. First, it allows the use of the same class of analytical

methods to study very diverse neural computations. In this context, a particularly important role is played by population-based classifiers (Bishop 2006). This type of approach can be used to measure the amount and format of information carried by neural ensembles encoding all sorts of information, including visual (Hung, Kreiman et al. 2005, Rust and DiCarlo 2010), auditory (Russ, Ackelson et al. 2008), olfactory (Shusterman, Smear et al. 2011) as well as more abstract concepts, such as numerosity (Tudusciuc and Nieder 2007).

Overview

In Chapter 2 we reveal a previously unknown untangling computation in PRH during visual target search. This computation acts to reformat inputs from IT, as suggested by the finding of increased linearly separable target match information in PRH, whereas the amount of total information was approximately matched in the two areas. Furthermore, we show that a simple linear-nonlinear model is sufficient to capture the transformation between the two brain regions.

In Chapter 3 we discuss the results of a set of analyses aiming at investigating the temporal dynamics of the representation of target match information in IT and PRH. Our finding was that signals first arrived in PRH in a tangled format, and were reformatted into an explicit format only after a delay of 10-15 ms. Surprisingly, it was possible to account for these dynamics using an extension of the model presented in

Chapter 1, and we further show that this was made possible by the non-stationarity of IT representation.

In Chapter 4 we provide an extensive description of a set of novel analytical tools, which were used to perform some of the analyses described in the previous chapters. In particular, we present a new method to design an orthonormal basis that can be used to decompose heterogeneous neural responses into a set of task-relevant, intuitive components. These components are shown to relate directly to measures of task-performance, such as the commonly used d' metric. In addition, methods are provided to prevent any bias in the estimation of such components.

Finally, in Chapter 5 we discuss how our results relate to the existing literature, and we speculate about possible future directions for this research.

CHAPTER 2: Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information

Marino Pagan, Luke S. Urban, Margot P. Wohl, and Nicole C. Rust (2013). Nature Neuroscience **16**(8): 1132–1139

Abstract

Theoretical models propose that the neural mechanisms responsible for finding visual targets include an initial combination of visual and target information, followed by a refinement process to determine whether a target is present in a currently-viewed scene. To investigate the specific neural computations responsible for the combination and refinement of visual and task-specific information, we recorded neural responses in inferotemporal cortex (IT) and perirhinal cortex (PRH) as macaque monkeys performed a task that required finding different targets in different blocks of trials. Our results suggest that visual and target information are initially combined within or before IT in a "tangled" or nonlinearly separable manner, followed by "untangling" computations in PRH that refine this information such that it is more accessible via a linear read-out. The neural computations responsible for untangling could be attributed to a mechanism that combines IT neurons with asymmetric tuning correlations.

Introduction

Searching for a specific object, such as your car keys, begins by activating and maintaining a representation of your target in working memory. Finding your target requires you to compare the visual content of a currently-viewed scene with this working memory representation to determine whether your target is currently in view. Our ability to rapidly and robustly switch between different targets suggests that this process is highly flexible. How does the brain achieve this?

Theoretical proposals of how the brain might find objects and switch between targets differ in their details (Salinas 2004, Sugase-Miyamoto, Liu et al. 2008, Salinas and Bentley 2009, Engel and Wang), but all propose that visual and target-specific working memory signals are first combined to produce a target-modulated visual representation, followed by second stage in which the combined signals are refined into a signal that reports when a currently-viewed scene contains a target (Fig 1). However, the means by which these signals are combined and refined remains little-understood. The initial combination of visual and working memory signals is likely to occur within higher stages of the ventral visual pathway (e.g. V4 and inferotemporal cortex, IT) via a process known as “feature-based” or “object-based” attention, as evidenced by V4 and IT neurons whose responses are modulated by both the identity of the visual stimulus as well as the identity of a sought target (Haenny, Maunsell et al. 1988, Maunsell, Sclar et al. 1991, Eskandar, Richmond et al. 1992, Lueschow, Miller et al. 1994, Gibson and Maunsell 1997, Liu and Richmond 2000, Chelazzi, Miller et al. 2001, Bichot, Rossi et al. 2005). While many models incorporate the simplifying assumption that the initial combination is implemented similarly by all neurons (e.g. a multiplicative enhancement of a visual tuning function), experimental evidence suggests that these initial mechanisms are in fact quite heterogeneous (Haenny, Maunsell et al. 1988, Maunsell,

Sclar et al. 1991, Eskandar, Richmond et al. 1992, Miller and Desimone 1994, Gibson and Maunsell 1997). These little-understood rules of combination likely determine the computations that the brain subsequently uses to determine whether a target is present in a currently-viewed scene.

To explore how visual and working memory signals are combined, we targeted IT, the highest stage of the ventral visual pathway. We find that this combination happens in a heterogeneous manner and one that results in a non-linearly separable or "tangled" (DiCarlo and Cox 2007) IT representation of whether a target is currently in view. To explore the computations by which this tangled representation is transformed into a report of whether a target is present, we recorded signals in PRH, which receives its primary input from IT (Suzuki and Amaral 1994) and has been demonstrated via lesioning studies to play a fundamental role in visual target search tasks (Meunier, Bachevalier et al. 1993, but see Buffalo, Ramus et al. 2000). Our results demonstrate that information about whether a target is currently in view is in fact "untangled" into a more linearly separable format in PRH and that the PRH population representation differs on correct as compared to error trials. Models fit to our data revealed that untangling is well-described by a linear-nonlinear mechanism that transforms nonlinearly separable information into a linearly separable format by combining signals from IT neurons that have a specific type of tuning correlations.

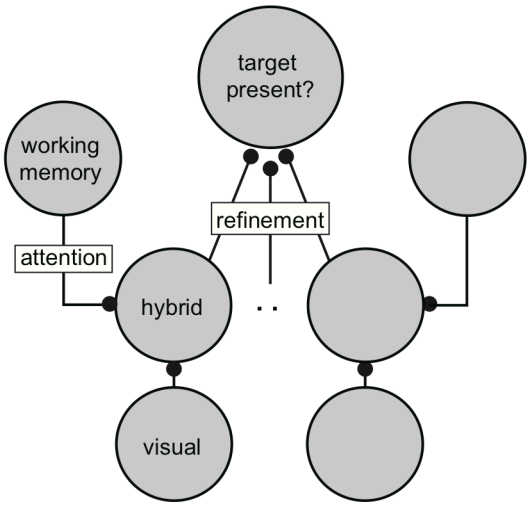

Figure 2-1. *Theoretical proposals of the neural mechanisms involved in finding visual targets*. Theoretical models propose that visual signals and working memory signals are nonlinearly combined in a distributed fashion across a population of neurons, followed by a refinement process to produce a representation that explicitly reports whether a target is present in a currently viewed scene. The initial nonlinear combination is thought to map onto the feature-based or object-based attention signals found at higher stages of the ventral visual pathway (i.e. V4 and IT), and thus the refinement process is likely to happen in association cortex (i.e. PRH and prefrontal cortex). The requirement for at least two stages of processing can be understood in the context that creating a signal that identifies whether a target is present, across changes in the identity of the target, falls under the class of “exclusive or” (XOR) problems (i.e. the solution requires a signal that identifies target matches as the conjunction of the visual stimulus N and a target object M where (N,M) can be (1,1) or (2,2) but not (1,2) nor (2,1)). Producing such a signal requires at least two stages of processing in a feed-forward network (Minsky and Papert 1969).

Methods

The subjects in this experiment were two adult male rhesus macaque monkeys (8.0 and 15.0 kg). Aseptic surgeries were performed to implant head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli, reward and data acquisition were controlled using customized software (http://mworks-project.org/). Stimuli were presented on a LCD monitor with a 85 Hz refresh (Samsung 2233RZ, (Wang and Nikolic 2011)). Both IT and PRH were accessed via a single recording chamber in each animal. Chamber placement was guided by anatomical magnetic resonance images and later verified physiologically by the locations and depths of gray and white matter transitions that included characteristic transitions through subcortical structures (e.g. the putamen and amydala) to reach PRH. The region of IT recorded was located on both the ventral superior temporal sulcus (STS) and the ventral surface of the brain, over a 4 mm medial-lateral region located lateral to the anterior middle temporal sulcus (AMTS) that spanned 14-17 mm anterior to the ear canals (Liu and Richmond 2000, Rust and DiCarlo 2010). The region of PRH recorded was located medial to the AMTS and lateral to the rhinal sulcus and extended over a 3 mm medial-lateral region located 19-22 mm anterior to the ear canals (Liu and Richmond 2000). Neural activity was recorded via a combination of glass-coated tungsten single electrodes (Alpha Omega, Inc.) and 16- and 24-channel U-probes with recording sites arranged linearly and separated by 150 micron spacing (Plexon Inc.). Continuous, wideband neural signals were amplified, digitized at 40kHz and stored via the OmniPlex

Data Acquisition System (Plexon, Inc.). All spike sorting was performed manually offline using commercially available software (Plexon, Inc.).

Details about the task can be found in Fig 2. Responses were only analyzed on correct trials, unless otherwise stated. Target matches that were presented after the maximal number of distractors (n=3) occurred with 100% probability and were discarded from the analysis. Unless otherwise stated, the response of each neuron was measured as the spike count in a time window 80 ms to 270 ms after stimulus onset. To maximize the length of our counting window but also ensure that spikes were only counted during periods of fixation, responses to target matches were selected from the 74.2% of correct trials on which the monkeys' reaction times exceeded 270 ms. Including trials with faster reaction times did not change the results reported here (i.e. claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys). As a measure of unit isolation, we determined the signal-to-noise ratio (SNR) of each waveform as the difference between the maximum and minimum of the mean waveform trace, divided by two times the standard deviation across the differences between the actual waveforms and the mean waveform (Kelly, Smith et al. 2007). Units were screened by their SNR and by a one-way ANOVA to determine those units whose firing rates were significantly modulated by the task parameters. When determining the screening criteria to include units in our analysis, we were concerned that setting any particular fixed value, particularly a highly stringent value, might differentially affect the two populations (e.g. due to lower firing rates in one of our populations). The most liberal screening procedure we applied (one-way ANOVA $p < 0.05$ and SNR > 2) resulted in 167 and 164 units in IT and PRH, respectively, and for all but the analysis shown in Fig 4e-f, these are the criteria we used for the Results. SNR was not statistically different in

the two resulting populations, as assessed by a statistical comparison of their means (mean IT = 3.47, PRH=3.55, p=0.55). Applying increasingly stringent criteria to the ANOVA (to $p<0.0001$) or to unit isolation (to SNR > 3.5) did not change the results (i.e. claims of significant and non-significant differences between IT and PRH for the data pooled across the two monkeys).

To assess the impact of simultaneous trial-by-trial variability (i.e. "noise correlations") on population performance (Fig 4d), we analyzed data simultaneously collected on the multi-channel U-probes (described above). During spike sorting, we defined at least one unit on every available channel, and we determined the 17 units from each session that produced the most significant p-values in the one-way ANOVA screen (without setting an absolute threshold on this p-value nor on SNR isolation). Linear classifier performance was assessed for these simultaneously recorded subpopulations in the manner described below. We used a similar approach to compute population performance on error trials (Fig 4f). Specifically, for each multi-channel recording session, we determined misses as instances in which the monkey failed to break fixation in response to the target match and false alarms as instances in which the monkey's eyes made a downward saccade in response to a distractor. We confined our analysis to false alarms in which the monkey fixated for a minimum of 270 ms before the response and for both types of error trials, we counted spikes in the same window used on correct trials (80 to 270 ms after stimulus onset). Linear classifier performance was compared on error and correct control trials in the manner described below.

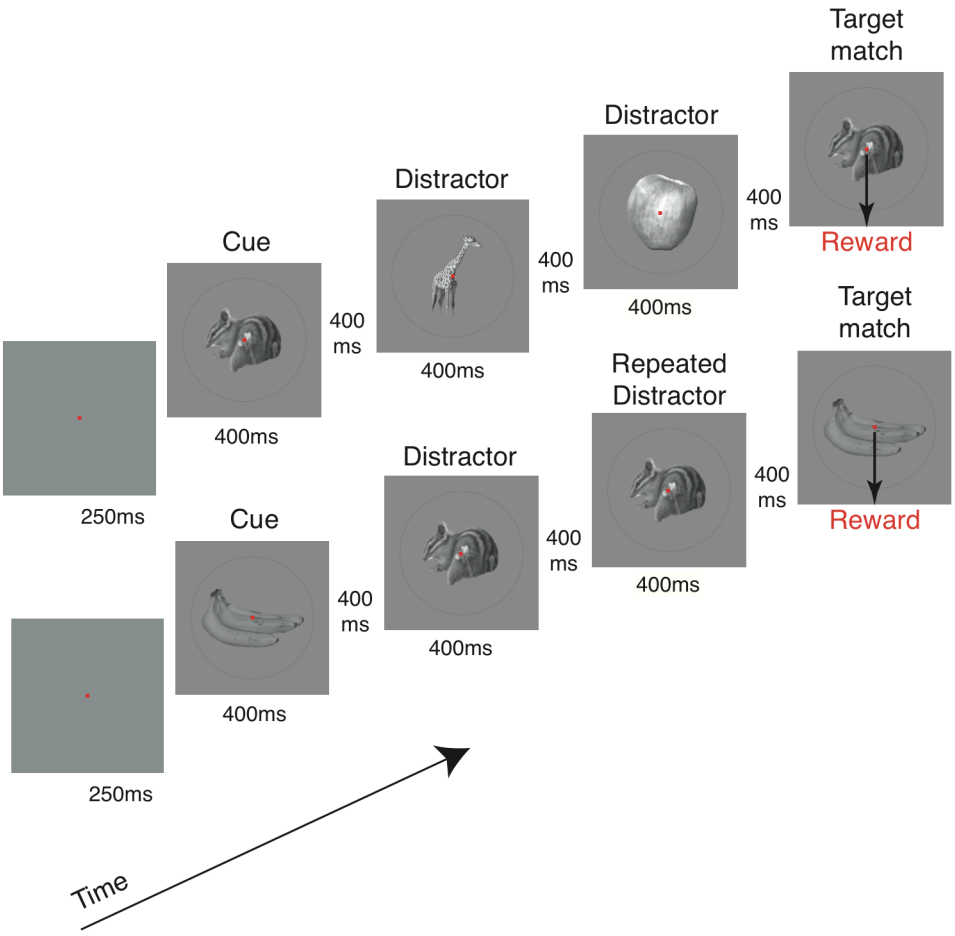

Figure 2-2. *The delayed-match-to-sample (DMS) task.* Monkeys were trained to perform a DMS task that required them to treat the same images as target matches and as distractors on different trials. Monkeys initiated a trial by fixating a small dot. After a 250 ms delay, an image indicating the target was presented, followed by a random number (0-3, uniformly distributed) of distractors, and then the target match. Each image was presented for 400 ms, followed by a 400 ms blank. Monkeys were required to maintain fixation throughout the distractors and make a saccade to a response dot located 7.5 degrees below fixation before the onset of the next stimulus to receive a reward. The same 4 images were used during all the experiments. Approximately 25% of trials included the repeated presentation of the same distractor with zero or one intervening distractors of a different identity. The same target remained fixed within short blocks of ~1.7 minutes.

Measures of the strength and congruency of visual and target modulations

To measure the response modulations within our recorded populations, we began by performing a two-way analysis of variance (ANOVA), by splitting each neuron's total response variability σ^2_{tot} (i.e. total variance across all trials and conditions) into four terms: modulation across visual stimuli σ^2_{vis}, modulation across sought targets σ^2_{targ}, nonlinear interactions of visual and target modulations σ^2_{NL}, and trial-by-trial variability σ^2_{noise}:

$$\sigma^2_{tot} \cdot \nu_{tot} = \sigma^2_{vis} \cdot \nu_{vis} + \sigma^2_{\text{targ}} \cdot \nu_{\text{targ}} + \sigma^2_{NL} \cdot \nu_{NL} + \sigma^2_{noise} \cdot \nu_{noise} \quad (1)$$

where $\nu_{tot} = 319$ (total number of degrees of freedom), $\nu_{vis} = 3$ (degrees of freedom of visual modulation), $\nu_{\text{targ}} = 3$ (degrees of freedom of target modulation), $\nu_{NL} = 9$ (degrees of freedom of visual/target modulation interactions), $\nu_{noise} = 304$ (degrees of freedom of noise variability). We then used the standard F-test to compare signal modulations to noise variability and establish which neurons had significant visual, target and/or nonlinear modulation.

Visual and cognitive modulation strength: The fraction of a neuron's variance that could be attributed to changes in the identity of the visual image (Fig 3b) was calculated as:

$$V_{strength} = \frac{\sigma^2_{vis}}{\sigma^2_{vis} + \sigma^2_{\text{targ}} + \sigma^2_{NL} + \sigma^2_{noise}} \quad (2)$$

The fraction of a neuron's variance that could be attributed to changes in the cognitive context (Fig 3c) was captured by the combined variance that could be attributed to linear and nonlinear target modulations (σ^2_{targ} and σ^2_{NL}):

$$C_{strength} = \frac{\sigma^2_{cog}}{\sigma^2_{vis} + \sigma^2_{cog} + \sigma^2_{noise}} \qquad \text{where} \qquad \sigma^2_{cog} = \sigma^2_{\text{targ}} + \sigma^2_{NL} \tag{3}$$

Because the term σ^2_{noise} combines both the actual trial-by-trial spiking variability as well as any trial-by-trial variability caused by recording noise, $V_{strength}$ and $C_{strength}$ can be regarded as noise-corrected measures of the proportions of visual and cognitive modulation. Both indices range from 0 to 1.

Congruency: For those neurons that were significantly modulated ($p<0.05$) by both visual and target information, or their interaction, we were interested in measuring the degree to which visual and target signals had been combined "congruently" (i.e. with similar object preferences). In doing so, it became necessary to evaluate congruency for the linear (σ^2_{vis} and σ^2_{targ}) and nonlinear interaction (σ^2_{NL}) terms separately. "Linear congruency" was defined as the absolute value of the Pearson correlation between the visual marginal tuning (i.e. the average response to each image as the visual stimulus) and the target marginal tuning (i.e. the average response to each image as the target):

$$lin\ congr = \left| \rho\left(x_{vis}, x_{\text{targ}} \right) \right| \tag{4}$$

$$x_{vis}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R\left(vis = i, \text{targ} = k \right) \qquad x_{\text{targ}}(i) = \frac{1}{4} \cdot \sum_{k=1}^{4} R(vis = k, \text{targ} = i)$$

where $R(vis = i, \text{targ} = k)$ is the average response to visual stimulus i, while searching for target k. To measure "nonlinear congruency", we considered the nonlinear modulation σ^2_{NL} described above and we sought to determine the degree to which these modulations fell along the diagonal (i.e. congruent nonlinear combinations of visual and target signals) versus off the diagonal (i.e. incongruent combinations). This was quantified by parsing the total nonlinear variability σ^2_{NL} into a term capturing the diagonal modulation σ^2_{diag} and a term capturing the non-diagonal modulation $\sigma^2_{nondiag}$:

$$\sigma^2_{diag} = \left(\mu_{Match} - \mu_{Distractor}\right)^2 / 3 \tag{5}$$

$$\sigma^2_{nondiag} \cdot \nu_{nondiag} = \sigma^2_{NL} \cdot \nu_{NL} - \sigma^2_{diag} \cdot \nu_{diag}$$

where $\nu_{NL} = 9$ (degrees of freedom of nonlinear interactions, as above), $\nu_{diag} = 1$ (degrees of freedom of diagonal modulation), $\nu_{nondiag} = 8$ (degrees of freedom of nondiagonal modulation). Nonlinear congruency was defined as the ratio between diagonal modulation and the sum of diagonal and nondiagonal modulation:

$$NL\ congr. = \frac{\sigma^2_{diag}}{\sigma^2_{diag} + \sigma^2_{nondiag}} \tag{6}$$

The final congruency index was computed as a weighted average of linear and nonlinear congruency, where the weights were determined by the firing rate variance for each term:

$$I_{congr} = \frac{\sigma^2_{lin} \cdot lin\ congr + \sigma^2_{NL} \cdot NL\ congr}{\sigma^2_{lin} + \sigma^2_{NL}} \quad \text{where} \quad \sigma^2_{lin} = \sigma^2_{vis} + \sigma^2_{\text{targ}} \tag{7}$$

This index ranges from 0 to 1.

Population performance

To determine population measures of the amount and format of information available in IT and PRH to discriminate target matches and distractors, we performed a series of classification analyses. Specifically, we considered the spike count responses of a population of N neurons to P presentations of M images as a population "response vector" **x** with a dimensionality equivalent to Nx1. We performed a series of cross-validated procedures in which (unless otherwise stated) we randomly assigned 80% of our trials (16 trials) to compute the representation ("training trials") and we set aside the remaining 20% of our data (4 trials) to test the representation ("test trials"). Two types of classifiers were tested:

Linear classification - SVM: To determine how well each population could discriminate target matches from distractors across changes in target identity using a linear decision rule, we implemented a linear readout procedure similar to that used by (Rust and DiCarlo 2010). The linear readout amounted to using the training data to find a linear hyperplane that would best separate the population response vectors corresponding to all of the target match conditions from the response vectors corresponding to distractors (Fig 4b, left). The linear readout took the following form:

$$f(x) = w^T x + b \tag{8}$$

where **w** is a Nx1 vector describing the linear weight applied to each neuron (and thus defines the orientation of the hyperplane), and **b** is a scalar value that offsets the hyperplane from the origin and acts as a threshold. The population classification of a test response vector was assigned to a target match when f(x) exceeded zero and was classified as a distractor otherwise. The hyperplane and threshold for each classifier was determined by a support vector machine (SVM) procedure using the LIBSVM library (http://www.csie.ntu.edu.tw/cjlin/libsvm) with a linear kernel, the C-SVC algorithm, and cost (C) set to 0.1.

Ideal observer classification: To determine how well each population could discriminate target matches from distractors across changes in target identity using an ideal observer, we computed the average firing rate response **r** of each neuron **u** to each of the 16 different conditions **c** from the training trials, r_{uc}. Assuming Poisson trial-by-trial variability, the likelihood that a test response **k** arose from a particular condition for a neuron was computed as the Poisson probability density:

$$lik_{u,c} = \frac{\left(r_{uc}\right)^{k} \cdot e^{-r_{uc}}}{k!} \tag{9}$$

We then computed the likelihood that a test response vector **x** arose from each condition **c** for the population as the product of the likelihoods for the individual neurons. Finally, we computed the likelihood that a test response vector arose from the category “target match” versus the category “distractor” as the mean of the likelihoods for target matches and distractors, respectively. The population classification was assigned to the category with the higher likelihood (Fig 4b, right).

To compare population performance between the different classifiers, we performed the same resampling procedure for each of them. On each iteration of the resampling, we randomly assigned trials without replacement for training and testing and when subpopulations with fewer than the full population were tested, we randomly selected a new subpopulation of neurons without replacement from all neurons. Because some of our neurons were recorded simultaneously but most of them were recorded in different sessions, unless otherwise stated, trials were shuffled on each iteration to destroy any (real or artificial) trial-by-trial correlation structure that might exist between neurons. Our experimental design resulted in 4 target match conditions and 12 distractor conditions; on each iteration we randomly selected 1 distractor condition from each image (for a total of 4 distractor conditions) to avoid artificial overestimations of classifier performance that could be produced by taking the prior distribution into account (e.g. scenarios in which the answer is more likely to be “distractor” than “target match”). Mean and error bars for performance were calculated as the mean and standard deviation, respectively, across 200 resampling iterations.

To assess the impact of correlated noise on population performance, we compared classifier performance when the trial-by-trial variability was kept intact as compared to when it was randomly shuffled (Fig 4d), for populations of 17 simultaneously recorded sites (where the data were extracted in the manner described above). Performance was computed as the mean across recording sessions; standard error was computed as the standard deviation across 200 iterations in which trials were randomly assigned as training and testing, for populations smaller than 17, the subset of neurons was randomly selected, and for the “shuffled noise” case, trials were randomly shuffled. To compare performance on correct and error trials (Fig 4f), we extracted the

error trials from these same multi-channel recording sessions. For each error trial (misses and false alarms; described above), we randomly selected a correct trial condition that was matched for the same target and visual stimulus as the condition that led to the error. We set aside these correct (and error) trials for cross validation, and trained the linear classifier on separate correct trials, as described above. Performance on each resampling iteration was computed as the average across all recording sessions; standard error was computed as the standard deviation across 800 resampling iterations in which correct trials were randomly assigned as training and test, and, for populations smaller than 17, the subset of neurons were randomly selected.

Modeling the transformation from IT to PRH

Single-neuron measure of linearly separable target information: As a single-neuron measure of match/distractor linear discriminability, we computed how well a neuron could linearly separate the responses to 4 target matches from the responses to 12 distractors (Fig 5a). This was measured by the squared difference between the mean response to all target matches μ_{Match} and the mean response to all distractors $\mu_{Distractor}$, divided by the variance of the spike count across trials, averaged across all 16 conditions σ^2_{noise} (Averbeck and Lee 2006):

$$I_L = \frac{\left(\mu_{Match} - \mu_{Distractor}\right)^2}{\sigma^2_{noise}} = \frac{\left(\mu_{Match} - \mu_{Distractor}\right)^2}{\mu} \quad (10)$$

Under the assumption that the responses are Poisson distributed, σ^2_{noise} can be replaced with the mean responses across all conditions, μ. Note that the numerator of this measure is proportional to the amount of diagonal structure in the matrix, as quantified by the measure σ^2_{diag} in Equation 5.

Static nonlinear model of the transformation from IT to PRH: To fit a nonlinear model (the "N model"; Fig 5c), the nonlinearity applied to each IT neuron Φ was defined as a monotonic piecewise linear function, with a threshold and saturation:

$$\Phi(x_i) = \begin{cases} k_{thr} & if\ x_i < k_{thr} \\ x_i & if\ k_{thr} \le x_i \le k_{sat} \\ k_{sat} & if\ x_i > k_{sat} \end{cases} \quad (11)$$

where k_{thr} indicates the threshold value, k_{sat} indicates the saturation value and x_i indicates the mean response of the IT neuron to condition i. Note that if k_{thr} is lower than x_i and k_{sat} is larger than x_i for all conditions then no nonlinearity is applied, so the formulation allows for the extreme case where $\Phi(x) = x$.

When applying this nonlinearity, we wished to avoid artificially creating information by applying transformations that could not be physically realized by neurons. Specifically, it is important to note that Linear-Nonlinear-Poisson (LNP) models operate by applying a nonlinearity to the mean neural responses across trials, and then simulate trial-by-trial variability with a Poisson process. In contrast, actual neurons can only operate on their inputs on individual trials, and thus their computations are influenced by

the trial-by-trial variability of their inputs. As an example, consider a toy neuron receiving only one input: when condition A is presented on three different trials the neuron receives 7, 8, and 9 spikes; when condition B is presented on three trials, the neurons receives 8, 9, and 10 spikes. The mean input is thus 8 spikes for condition A and 9 spikes for condition B. An LNP model might attempt to take these inputs and apply a threshold at 8.5 spikes, below which it might set the firing rate to 0 spikes; such a nonlinearity would set the mean response to 0 spikes for condition A and 9 spikes for condition B, and after Poisson noise was regenerated, the distribution of responses for conditions A and B would be highly nonoverlapping (e.g. Poisson draws for condition A might be 0, 0, 0 and Poisson draws for condition B might be 8, 9, and 10). However, artificially separating the input distributions in this way by a threshold violates laws of information processing. This can be demonstrated by noting that if the same threshold were applied trial-by-trial, it would produce 0, 0 and 9 spikes for condition A (mean 3) and 0, 9, and 10 spikes for condition B (mean 6.3), thus preserving the fact that the two distributions are in fact overlapping. In our model we aimed at exploiting the simplicity and expressive power of LNP models while also taking trial-by-trial response variability into consideration such that we did not artificially create information. Our strategy was twofold: first, we constrained the model by imposing that nonlinearities could only reduce the difference between the means of any pair of conditions. This was accomplished by imposing that matrix values could only be “squashed” towards the threshold and the saturation, i.e. values below the threshold are set to the value of threshold, and values above the saturation are set to the saturation value (see Equation 11). Second, we renormalized the response matrix after applying the nonlinearity to ensure that the overall signal-to-noise ratio was not artificially increased by the generation of Poisson

variability. In particular, we made the conservative assumption that the trial-by-trial variability was not modified by the nonlinearity, and therefore was equal to the mean response across all conditions before the application of the nonlinearity μ_{before} (see Equation 10). If the overall mean response was shifted by the nonlinearity to a new value μ_{after}, it was necessary to rescale the matrix to insure that the signal to noise ratio was consistent with the true variability, equal to μ_{before} (i.e. no information was artificially created). This was accomplished by multiplying the response matrix by the ratio of μ_{after} and μ_{before}:

$$M_{normalized} = M \cdot \frac{\mu_{after}}{\mu_{before}} \tag{12}$$

where M indicates the response matrix before normalization, and $M_{normalized}$ is the response matrix after normalization.

When fitting the N model to our data (Fig 5c), we explored all possible nonlinearities by allowing k_{thr} and k_{sat} to take any of the values in the original response matrix, for a total of 120 possible nonlinearities. The selected values were those that maximized the linearly separable target information (I_L, Equation 10).

Pairwise linear-nonlinear model of the transformation from IT to PRH: Pairs of model PRH neurons were created via two orthonormal linear combinations of pairs of IT neurons, each followed by a static monotonic nonlinearity, that maximized the joint linearly separable information of the two model PRH neurons. Here we define the response matrices of the two "input" IT cells as I_1 and I_2; the response matrices of the

two “output” neurons as O_1 and O_2 ; the weights of the two linear combinations (indexed by input neuron, output pair) as w_{11} , w_{21} , w_{12} and w_{22} ; and the two monotonic nonlinearities as Φ_1 and Φ_2 .

$$O_1 = \Phi_1\left(w_{11} \cdot I_1 + w_{12} \cdot I_2\right) \quad ; \quad O_2 = \Phi_2\left(w_{21} \cdot I_1 + w_{22} \cdot I_2\right) \tag{13}$$

where orthogonality of the weights was imposed by:

$$w_{11} \cdot w_{21} + w_{12} \cdot w_{22} = 0 \tag{14}$$

and each pair of weights was constrained to a unitary norm:

$$w_{11}^2 + w_{12}^2 = 1 \quad ; \quad w_{21}^2 + w_{22}^2 = 1 \tag{15}$$

Because the weights were orthogonal and each pair was constrained to be unit norm, the weights could be defined as a rotation matrix:

$$W = \begin{bmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{16}$$

where θ is the angle by which the two-dimensional response space is rotated around the origin by the linear operation (compare Fig. 6c, bottom left versus center). Constraining the weights to be orthonormal is both necessary and sufficient to insure that no information is copied in the newly-created neurons: the original space is simply rotated and the separation between the response clouds to different conditions are left intact. Conversely, non-orthogonal weights would result in “copying over” the original information multiple times (note that copying the original information multiple times would not lead to an overall increase of the total information because the trial-by-trial variability

in the two newly created neurons would be correlated). To find the optimal linear combinations for each pair of IT cells, we exhaustively explored all possible angles by systematically varying θ from 1 to 360 degrees. When responses were negative (i.e. as a result of negative weights), the values of the response matrix were shifted to positive values and the matrix was renormalized to ensure that the shifting process did not artificially create information. This procedure resembles the renormalization we applied for static nonlinearities (Equation 12). First we estimated the average trial-by-trial variability in the output matrix as the weighted combination of the average noise variances of the two input neurons:

$$\sigma_O^2 = w_1^2 \cdot \sigma_{I1}^2 + w_2^2 \cdot \sigma_{I2}^2 \tag{17}$$

where σ_O^2 is the noise variability in the output neuron, w_1 and w_2 are the weights, and σ_{I1}^2 and σ_{I2}^2 are the noise variances of the two input neurons. Next, we normalized the shifted response matrix $M_{shifted}$ by multiplying it by the ratio between its mean response $\mu_{shifted}$, and the actual predicted output noise σ_O^2:

$$M_{normalized} = M_{shifted} \cdot \frac{\mu_{shifted}}{\sigma_O^2} \tag{18}$$

This ensured that the overall signal-to-noise ratio could not be influenced by changes in the mean response (i.e. average noise variance under the Poisson assumption) due to the nonlinearity or the shift required to make all response values non-negative.

When considering our input population, we allowed for “shifted copies” of our recorded IT neurons. More specifically, we allowed the model to make one selection

from the set defined by each actual IT matrix we recorded and the 23 permutations of that matrix that are obtained by simultaneously shifting the four rows and four columns of the matrix. This procedure preserves the rules of combination between visual and working memory information (i.e. the strengths of visual and cognitive modulation and their congruency; Fig 3) but shifts their object preferences. Stated differently, our assumption is that the rules of combination of visual and working memory signals are not specific to the object preferences of a neuron (i.e. the brain does not employ one rule of combination for apple preferring neurons and a different rule for banana preferring neurons) and that any inhomogeneities with regard to object preferences that are included in our data set (e.g. an excess of selective match detectors for object 1 as compared to object 4) are due to finite sampling. For every possible pair of IT neurons, we generated all possible output neurons by considering all 24 matrix permutations, each paired by 360 possible angles, and each of those with all 120 possible nonlinearities. We also searched similar parameters for all possible pairs of output neurons generated by orthogonal weights to determine the pairing parameters that produced maximal joint linearly separable information.

Having determined the best parameters for every possible pair of IT neurons, we selected the subset of pairings that produced a model PRH population with the maximal amount of total linearly separable information while only allowing each IT input neuron to contribute to the model output population once. This selection problem can be reduced to an integer linear programming problem (Edmonds and Johnson 1970), and we implemented a standard solution using the GLPK library (http://www.gnu.org/software/glpk).

Untangling via asymmetric match and distractor tuning correlations

Upon establishing that the pairwise LN model was effective at transforming nonlinearly separable information into a linearly separable format (Fig 5), we were interested in an intuitive (and yet quantitatively accurate) understanding how the model worked. Given any neuron's response matrix, one crucial property that enables a monotonic nonlinearity to extract linearly separable information (i.e. to increase the distance between the mean response to the matches and the mean response to the distractors) is the degree to which the "tails" of the match and distractor distributions are non-overlapping (Fig. 6a). Although one could, in theory, fully characterize the match and distractor distributions and arrive to a closed-form estimate of the maximum extractable linearly separable information in a neuron's matrix via a nonlinearity, we focused on producing a simple estimate of this quantity based just on the first two moments of these distributions (i.e. their means and variances). We postulated that the absolute value of the difference in variance across the matches (σ^2_{Match}) and the variance across the distribution of distractors ($\sigma^2_{Distractor}$) is a good predictor of the amount of linearly separable information that can be extracted by a monotonic nonlinearity (Δ_{info}):

$$\Delta_{\text{info}} \approx k \cdot \left| \sigma^2_{Match} - \sigma^2_{Distractor} \right| \tag{19}$$

where k is a proportionality constant. This estimate assumes that the means of the match and distractor distributions are the same and that variance differences thus

translate into regions in which the high-variance distribution extends beyond the low-variance distribution (Fig 6a). An improvement of this estimate can be obtained by correcting for the fact that the initial distance between the means of the two distributions (i.e. the amount of pre-existing linearly separable information) always decreases the amount of overlap and thus always limits the amount of further information that can be extracted (see Fig. 6b):

$$\Delta_{\text{info}} \approx k \cdot \max\left(0, \Delta\sigma^2 - \left(\Delta\mu\right)^2\right) \tag{20}$$

To extend the prediction to pairs of neurons, one must consider the covariance matrix for the bivariate distribution of match responses Σ_{Match} and of distractor responses $\Sigma_{Distractor}$, which can be further decomposed into the variances across matches and distractors and on the tuning correlations for matches and distractors between the two neurons. Because the amount of linearly separable information gained by a pairing is proportional to the absolute value of the difference of the variances for matches and distractors ($\Delta\sigma^2$, Equation 19), the model will tend to pair IT neurons that maximize $\Delta\sigma^2$. Here we derive the amount of $\Delta\sigma^2$ that results from a pairing. First, the variance across match responses $\sigma^2_{Match,lin.comb.}$ for a linear combination with weights w_1 and w_2 can be computed as:

$$\sigma^2_{Match,lin.comb.} = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \cdot \Sigma_{Match} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \quad \ldots$$

$$= \begin{bmatrix} w_1 & w_2 \end{bmatrix} \cdot \begin{bmatrix} \sigma^2_{Match,1} & \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} \\ \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} & \sigma^2_{Match,2} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \quad \ldots$$

$$= w_1^2 \cdot \sigma^2_{Match,1} + w_2^2 \cdot \sigma^2_{Match,2} + 2 \cdot w_1 \cdot w_2 \cdot \rho_{Match} \cdot \sigma_{Match,1} \cdot \sigma_{Match,2} \quad (21)$$

Analogously, the variance across distractors for the linear combination $\sigma^2_{Distractor,lin.comb.}$ can be computed as:

$$\sigma^2_{Distractor,lin.comb.} = w_1^2 \cdot \sigma^2_{Distractor,1} + w_2^2 \cdot \sigma^2_{Distractor,2} + 2 \cdot w_1 \cdot w_2 \cdot \rho_{Distractor} \cdot \sigma_{Distractor,1} \cdot \sigma_{Distractor,2} \quad (22)$$

Consequently, the difference between variances can be obtained by subtracting (22) from (21):

$$\Delta\sigma^2_{lin.comb.} = w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 + 2 \cdot w_1 \cdot w_2 \cdot \left(\rho_{Match} \cdot \bar{\sigma}^2_{Match} - \rho_{Distractor} \cdot \bar{\sigma}^2_{Distractor} \right) \quad (23)$$

where $\Delta\sigma_1^2$ indicates the match/distractor variance difference for input neuron 1, $\Delta\sigma_2^2$ indicates the variance difference for input neuron 2, $\bar{\sigma}^2_{Match}$ is the geometric mean of the variances for matches of the two neurons, and $\bar{\sigma}^2_{Distractor}$ is the geometric mean of the variances for distractors. It is evident from equation 23 that variance difference between matches and distractors after pairing can derive from two different sources. First, variance differences can be inherited from the input neurons ($\Delta\sigma_1^2$ and $\Delta\sigma_2^2$):

$$\Delta\sigma^2_{lin.comb.} \approx w_1^2 \cdot \Delta\sigma_1^2 + w_2^2 \cdot \Delta\sigma_2^2 \quad (24)$$

For this type of variance difference, pairing is not required as linearly separable information could be extracted by applying a nonlinearity to each of the input matrices individually (Fig 6a). Second, variance differences that did not exist in the inputs can be produced via asymmetric tuning correlations for matches and distractors:

$$\Delta\sigma^2_{lin.comb.} \approx 2 \cdot w_1 \cdot w_2 \cdot \left(\rho_{Match} \cdot \bar{\sigma}^2_{Match} - \rho_{Distractor} \cdot \bar{\sigma}^2_{Distractor} \right) \quad (25)$$

As demonstrated in Fig 5c, the ability of the pairwise LN model to extract linearly separable information relies heavily on this second source of variance difference (compare the N model to the LN model). Finally, a prediction of how these variance differences translate into increases in linearly separable information can be made by applying equation 20 with the empirically derived constant of $k = 0.15$ applied to all pairs. Despite the great simplicity of this description and the fact that only the first two moments (mean, variance and covariance) of the match and distractor distributions are considered, this estimate is quite reliable at predicting the gain in linearly separable information in the model (Pearson correlation between the increase in linearly separable information for each LN model pair and the prediction (Equation 23): $r = 0.84$, $r^2 = 0.7$).

Statistical tests

For each of our single neuron measures, we report p-values as an evaluation of the probability that differences in the mean values that we observed in IT versus PRH were due to chance. As many of these measures were not normally distributed, we calculated these p-values via a bootstrap procedure (Efron and Tibshirani 1994). On each iteration of the bootstrap, we randomly sampled the true values from each population, with replacement, and we computed the difference between the means of the two newly created populations. The p-value was computed as the fraction of 1000 iterations on which the difference was flipped in sign relative to the actual difference between the means of the full dataset (e.g. if the mean for PRH was larger than the

mean for IT, the fraction of bootstrap iterations in which the IT mean was larger than the PRH mean).

Results

IT responses reflect heterogeneous mixtures of visual and target information

We recorded neural responses in IT and PRH as monkeys performed a delayed-match-to-sample, sequential object search task that required treating the same images as targets and as distractors on different trials (Fig 2). Behavioral performance was high overall (monkey 1: 94% correct; monkey 2: 92% correct). Performance remained high on trials that included the same distractor presented repeatedly before the target match (monkey 1: 91% correct; monkey 2: 87% correct), confirming that the monkeys were generally looking for specific images as opposed to detecting the repeated presentation of any image (consistent with Miller and Desimone 1994). Altogether, four images were presented in all possible combinations as a visual stimulus ("looking at"), and as a target ("looking for"), resulting in a four-by-four response matrix (Fig 3a). As monkeys performed this task, we recorded neural responses in IT and PRH. To examine response properties, unless otherwise stated, we counted spikes after the onset of each test (i.e. non-cue) stimulus within a window that accounted for neural latency but also preceded the monkeys' reaction times (80 - 270 ms). We then screened for neurons that were significantly modulated across the 16 conditions, as assessed by a one-way ANOVA (see Methods). Only the data from correct trials are reported here.

We note that the three components of this task (described above) each produce distinct structure in these response matrices: “visual” selectivity translates to vertical structure, “working memory” selectivity for the current target translates to horizontal structure, and because matches fall along the diagonal of this matrix and distractors fall off the diagonal, differential responses to target matches and distractors translates to diagonal structure (Fig 3a). To characterize the responses of both populations, we began by parsing each neuron’s matrix into the fraction of firing rate modulations that could be attributed to: 1) changing the visual image (Fig 3b), 2) changing the cognitive context (i.e. the combined modulations that could be attributed to changing the identity of the target and/or whether the condition was a match or a distractor, Fig 3c), and 3) noise due to trial-by-trial variability (see Methods, Equations 1-3). On average, both populations were strongly visually modulated but also had cognitive modulations that were well above the noise (means IT: visual=74%, cognitive=21%, noise=5%; PRH: visual=62%, cognitive=31%, noise=7%; IT vs. PRH mean comparisons: visual $p<0.0001$; cognitive $p<0.0001$). For neurons whose matrices reflected mixtures of visual and cognitive information, we were interested in understanding the degree to which visual and working memory information were combined “congruently” with similar object preferences for visual stimuli and targets (e.g. the "Selective target detector" in Fig 3a responds with a high firing rate to object 4 when presented as both a visual stimulus and as a target; see also Maunsell, Sclar et al. 1991 Fig 12) as opposed to “incongruently”, with different object preferences (e.g. the "distractor detector" in Fig 3a responds with a high firing rate when object 2 is presented as a visual stimulus and object 3 is the target; see also Maunsell, Sclar et al. 1991 Fig 13). We developed a metric for congruency (see Methods, Equation 7) and found that in IT, highly congruent and incongruent

combinations were nearly equally likely, with a slight overall bias toward congruency (Fig 3d, gray); whereas in PRH, congruency measures were shifted toward slightly higher values (Fig 3d, black outline; means comparison p=0.01). The existence of incongruent neurons could not be explained by neurons with poor single-unit isolation (Pearson correlation of the signal-to-noise ratio (SNR) measure of isolation and congruency: IT r=0.04; PRH r=-0.001). In sum, our results suggest that visual and working memory signals are reflected heterogeneously in the response matrices of IT and PRH neurons with regard to their relative strengths and the congruency of their combination.

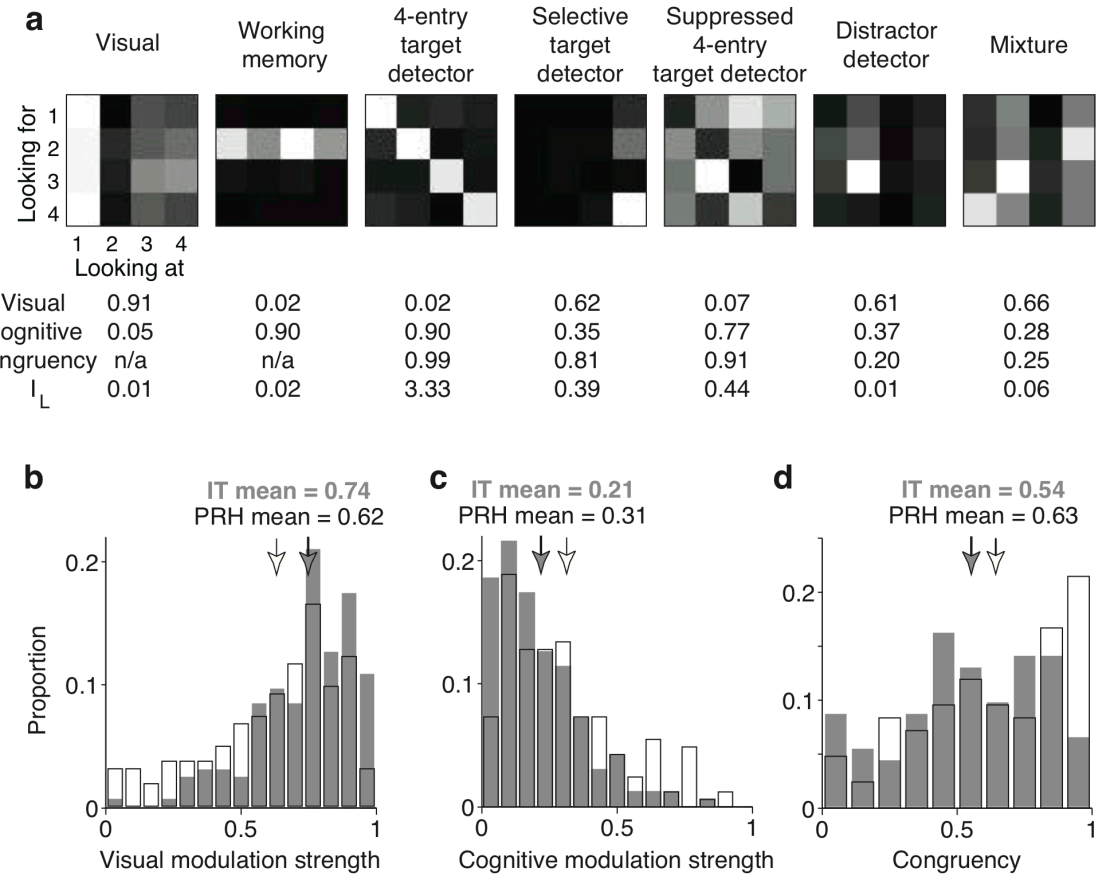

Figure 2-3. *Example responses.* **a)** Each of four images were presented in all possible combinations as a visual stimulus ("looking at"), and as a target ("looking for"), resulting in a four-by-four response matrix. Shown are the response matrices for example neurons with different types of structure (labeled). Some neurons reflected nearly pure versions of the task parameters (e.g. "Visual", "Working memory", "4-entry target detector"). Other neurons reflected conceptually intuitive variations of those parameters (e.g. the "Selective target detector" responds to one object as a target match and the "Suppressed 4-entry target detector" reflects target matches with decreases in firing rate; see also Miller and Desimone 1994). The responses of other neurons were more difficult to interpret (e.g. the "Distractor detector" and the "Mixture" neuron). All matrices depict a neuron's response with pixel intensity proportional to firing rate, normalized to range from black (the minimum) to white (the maximum) firing rate. Also labeled: single-neuron measures of "Visual": the strength of visual modulation (subpanel b), "Cognitive": the strength of cognitive modulation (subpanel c), "Congruency" the alignment of visual and target signals (subpanel d), "I_L": single-neuron linearly separable information (described under the section labeled "A pairwise linear-nonlinear model can account for PRH untangling"). **b-d)** Histograms of the IT (gray) and PRH (black outline) populations for various measures of response modulation. All metrics have bounds of 0-1. Arrows indicate means. To calculate these measures, each neuron's matrix was first parsed into firing rate modulations that could be attributed to: visual, cognitive, and noise (i.e. trial-by-trial variability; see Methods). **b)** The proportion of each neuron's response modulations that could be attributed to changing the identity of the visual stimulus. **c)** The proportion of each neuron's response modulations that could be attributed to changing the cognitive context, which is defined as changing the target identity and/or whether the condition was a target match or a distractor. **d)** "Congruency", a measure of the degree to which each neuron's responses reflected target modulations that were aligned to its preferences for visual stimuli, where 1 indicates perfect alignment and 0 indicates complete misalignment (see Methods). Only the 93/167 IT and 84/164 PRH neurons that met the criteria for calculating congruency (i.e. were either significantly modulated by both visual and working memory signals or by their interaction) are included.

Target match information is "untangled" between IT and PRH

How do the heterogeneous responses of IT and PRH neurons relate to a determination of whether a currently-viewed image matches the sought target (i.e. the solution to the monkey's task)? To assess this, we began by probing target match/distractor information in the IT and PRH populations with a linear-readout (Fig 4a, right). More specifically, we determined how well a linear decision boundary could separate target matches from distractors via a cross-validated analysis that involved using a subset of the data to find the linear decision boundary via a machine learning procedure (SVM) and we then tested the boundary with separately measured trials (see Methods; Equation 8). Cross-validated population performance was significantly higher in PRH than in IT (Fig 4b, left) and this was confirmed in each monkey individually (Fig 4c, "M1" and "M2"). Higher PRH performance could not be explained by the repeated presentation of the "match" after it had previously been presented in the trial as the "cue" (Fig 4c, "Adaptation control") nor by changes in reward expectation as a function of the number of distractors encountered thus far in a trial (Fig 4c, "Expectation control"). Finally, while the analyses described thus far assume trial-by-trial independence between neurons, correlated variability has been shown to impact linear read-out population performance for some tasks (Cohen and Maunsell 2009, Graf, Kohn et al. 2011). For our data, we tested the independence assumption by analyzing smaller subpopulations of simultaneously recorded neurons, and found similar results when the noise correlations were kept intact and when they were scrambled (Fig. 4d).

We were also interested in determining whether our recorded responses were consistent with a putative role in the circuitry that transforms sensory information into a behavioral response. Consistent with this hypothesis, PRH linear classification performance peaked well before the monkeys' reaction times, which were longer than 270 ms on these trials (Fig 4e). We also found that linear classification performance on error as compared to correct trials trended toward lower values in IT and was significantly lower in PRH (Fig 4f). Poorer error trial performance could not be attributed simply to a difference in firing rate (grand mean firing rates: IT correct = 7.6 Hz, error 7.2 Hz, p=0.26; PRH correct = 5.6 Hz; error 5.5 Hz, p=0.45).

The results we have presented thus far could arise from a scenario in which visual and working memory signals are combined within or before IT in a "tangled" or nonlinearly separable manner, followed by "untangling" computations in PRH that transform the IT input into a more linearly separable format. Alternatively, these results could arise from a scenario in which PRH receives its input not only from IT but also from one or more additional sources (e.g. via a separate working memory input from prefrontal cortex), thus resulting in more total information for this task in PRH. To discern between these alternatives, we probed the total information for this task in a manner that did not depend on the specific format of that information. More specifically, total information depends only the degree to which the response clouds corresponding to the sixteen different task conditions are separated (i.e. non-overlapping), but not on the specific manner in which the response clouds corresponding to each condition are positioned relative to one another (Fig 4a, compare center and right). As a measure of the total information available for match/distractor discrimination in the IT and PRH populations, we performed a cross-validated, ideal observer match/distractor

classification of the population response on individual trials (see Methods, Equation 9). Performance on this task was slightly lower in IT, but not significantly so (Fig 4b, right), suggesting that IT and PRH contain similar amounts of total information for this target-switching task. These results are consistent with a model in which PRH receives all or nearly all its information for this task from IT, as opposed to other sources.

Taken together, the results reported in Fig 4 are consistent with a model in which visual and working memory signals are initially combined within or before IT in the ventral visual pathway in a heterogeneous and "tangled" manner, followed by reformatting operations in PRH that "untangle" target match information. These results are reminiscent of the untangling phenomena described at earlier stages of the ventral visual pathway (i.e. from V1 to V4 to IT) for invariant object recognition (Hung, Kreiman et al. 2005, DiCarlo and Cox 2007, Rust and DiCarlo 2010), and thus demonstrate that the untangling applies not only to perceptual processing (i.e. identifying the content of a currently-viewed scene), but also extends to task-specific cognitive processing (i.e. find a specific sought target object).

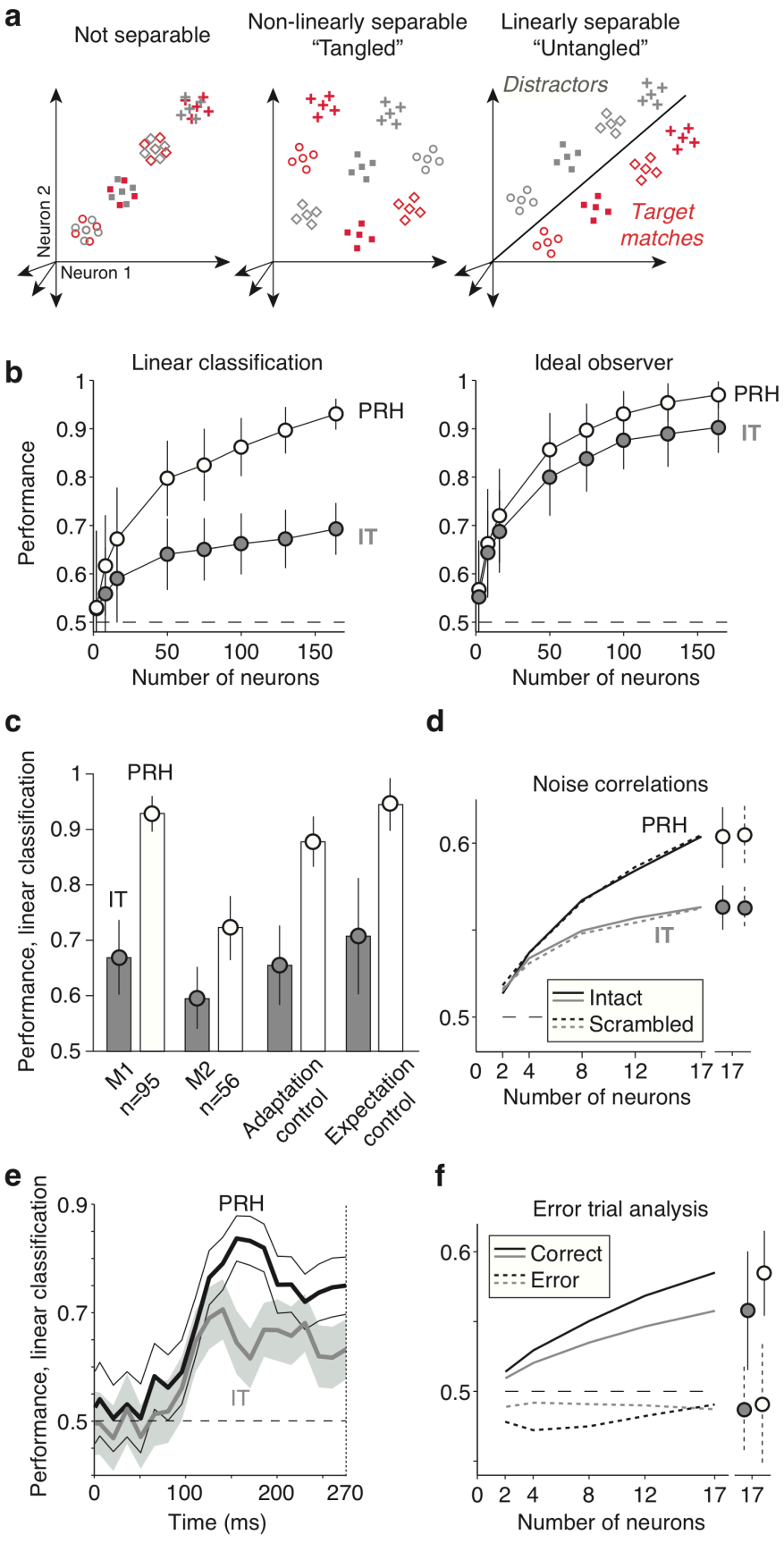
a
Not separable
Non-linearly separable
"Tangled"
Linearly separable
"Untangled"
Distractors
Target matches
Neuron 2
Neuron 1
b
Linear classification
Ideal observer
Performance
PRH
IT
Number of neurons
c
Performance, linear classification
M1 n=95
M2 n=56
Adaptation control
Expectation control
d
Noise correlations
Intact
Scrambled
e
Time (ms)
f
Error trial analysis
Correct
Error

Figure 2-4. *Population performance.* **a)** The relationship between different types of match/distractor separability and information. Each point depicts a hypothetical population response, consisting of a vector of the spike count responses to a single condition on a single trial. The four different shapes depict the hypothetical responses to the four different images and the two colors (red, gray) depict the hypothetical responses to target matches and distractors, respectively. For simplicity, only 4 of the 12 possible distractors are depicted. Clouds of points depict the predicted dispersion across repeated presentations of the same condition due to trial-by-trial variability. The target-switching task (Figure 2) requires discriminating the same objects presented as target matches and as distractors. Hypothetical scenarios include: no separability (left), non-linear separability (center), and linear separability (right). **b)** Performance of the IT (gray) and PRH (white) populations, plotted as a function of the number of neurons included in each population, via an analysis designed to probe linear separability (left), and total separability (linear and/or nonlinear; right). The dashed line indicates chance performance. Linear separability was probed with a cross-validated analysis that determined how well a linear decision boundary could separate target matches and distractors (see Text, Methods). Total separability was probed with a cross-validated, ideal observer analysis (see Text, Methods). **c)** Linear separability for the IT (gray) and PRH (white) populations plotted for: monkey 1 (M1) and monkey 2 (M2) individually; as a control for repeated stimulus effects (e.g. adaptation), performance for discriminating target matches and repeated distractors ("Adaptation control"); as a control for reward expectation as a function of the number of distractors encountered thus far in a trial and other position effects, performance for discriminating target matches and distractors presented as the first test stimulus after the cue ("Expectation control"). Note that due to a more limited number of trials available for the selection, spikes were counted in a shorter window (80 - 215 as compared to 80 - 270 ms) for the position-matched expectation control as compared to the other analyses**. d)** Comparison of linear classification performance for simultaneously recorded populations of size 17 (see Methods), when noise correlations (trial-by-trial variability) were left intact (solid) or scrambled (dotted). Shown is the average performance across 14 recording sessions in IT and 25 sessions in PRH. **e)** Evolution of population linear separability over time. Thick lines indicate performance of the entire IT (gray) and PRH (black) populations for

counting windows of 30 ms with 15 ms shifts between neighboring windows. Thin lines indicate standard error. The dotted line indicates the minimum reaction time on these trials (270 ms). **f)** Linear classification performance on error (dotted) as compared to correct (solid) trials. Each error trial was matched with a randomly selected correct trial that had the same target and visual stimulus as the condition that resulted in the error and both sets of trials were set aside from the training set to measure cross-validated performance when the population read-out was trained on correct trials, as described above. Error trials included both misses (of target matches) and false alarms (i.e. responding to a distractor). As in subpanel d, the analysis was performed for all multi-channel recording sessions and the average performance was computed across sessions. In all plots, dashed line indicates chance performance; error bars correspond to the variability that can be attributed to trials randomly assigned for training and testing, and for populations smaller than the full data set, the neurons randomly selected.

A pairwise linear-nonlinear model can account for PRH untangling

Having established that information is reformatted or "untangled" between IT and PRH, we were interested in understanding the neural mechanisms by which the responses of IT neurons were transformed into the responses of PRH cells. To determine these computations, we began by formulating a single-neuron measure of how much linearly separable or "untangled" information was conveyed by an individual neuron's response matrix. Linearly separable information depends on the separation of the responses to the match conditions and the responses to the distractor conditions (Fig 5a), and it can be calculated by the squared difference in the mean response to matches and the mean response to distractors, divided by the pooled trial-by-trial variability (see Methods, Equation 10). We note that the difference in the mean

response to matches and distractors maps directly onto the amount of "diagonal structure" in a neuron's response matrix (see Methods). Thus a "4-entry target detector" will have high linearly separable target match information, a "selective target detector" will have a bit less, and a highly visual neuron or working memory neuron will have almost none (see the example neurons in Fig 3a, where "I_L" values are labeled). Consistent with the population results presented in Fig 4b (left), we find that PRH has significantly higher single-neuron linearly separable target match information than does IT (mean IT=0.07; mean PRH=0.17, $p<0.0001$).

We then set out to determine the simplest class of models that could take our recorded IT responses as input and produce a model population that behaved like our recorded PRH. We began by ruling out a *priori* the class of models in which IT neurons combine linearly to produce PRH cells because we know that linear operations can move linearly separable information around within a population (i.e. between neurons) but cannot transform non-linearly separable information into a linearly separable format. Thus we began by testing the class of nonlinear models in which a static nonlinearity (i.e. thresholding and saturation; see Methods, Equation 11) was fit to each IT neuron such that its response matrix conveyed maximal linearly separable target match information. Inconsistent with the large gains in linear readout performance we observed from IT to PRH, we found only modest overall gains in this model population (Fig 5c, right, "N Model"). Next we considered the class of models in which pairs of IT neurons combine via a linear-nonlinear model ("LN model") to produce the responses of pairs of PRH cells (Fig 5b). In fitting our model, we imposed the important constraint that information could not be replicated multiple times in the transformation from IT to PRH (i.e. the same neuron could not be copied multiple times). To enforce this rule, our model

created two PRH neurons by applying two sets of orthonormal linear weights to the pair of IT inputs (e.g. $(+\sqrt{0.5},+\sqrt{0.5})$ and $(+\sqrt{0.5},-\sqrt{0.5})$) and each IT neuron was included only once (see Methods, Equations 13-15). We searched all possible pairwise combinations of IT neurons and nonlinearities and selected the combinations that produced the largest gains in linearly separable information (see Methods). The resulting LN model population nearly matched the population performance increases in PRH over IT with a linear readout and replicated PRH population performance on the match/distractor task with an ideal observer read-out (Fig 5c, "LN Model"). The LN model also replicated the single-neuron response properties reported in Fig 3, in terms of significantly different mean values than IT but not PRH (mean LN model visual modulation strength=0.65, vs. IT $p<0.0001$, vs. PRH $p=0.072$; mean LN model cognitive modulation strength=0.28, vs. IT $p<0.0001$, vs. PRH $p=0.113$; mean LN model congruency=0.63, vs. IT $p=0.006$, vs. PRH $p=0.45$). The fact that such a simple model can reproduce the transformation we observed in our data from IT to PRH provides additional support for the plausibility that PRH receives its inputs for this task directly from IT, as opposed to other sources. The simplicity of the model also lends itself to an exploration of the specific computational mechanisms underlying untangling, as described below.

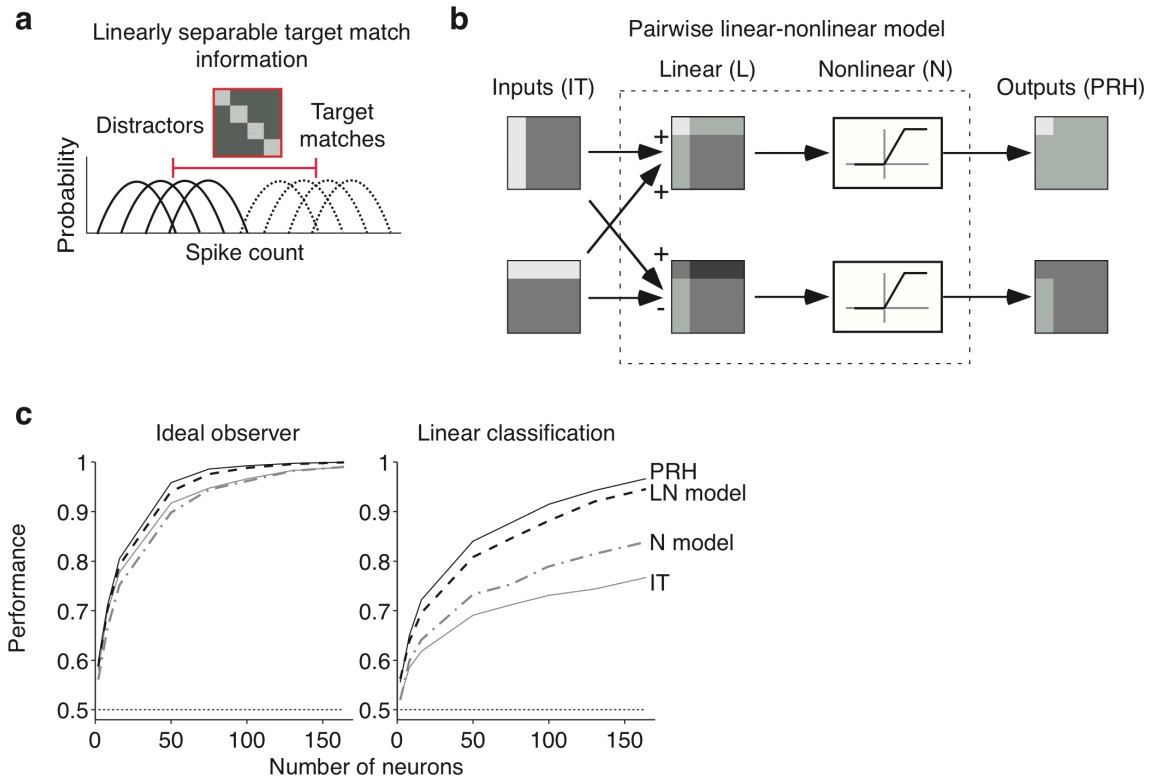

Figure 2-5. *Modeling the transformation from IT to PRH.* **a)** The amount of linearly separable target match information in a neuron's matrix relates to the separation in its responses to target matches (dashed) and distractors (solid lines). It is calculated as the squared difference of the average response to matches and the average response to distractors, divided by their pooled trial-by-trial variability (see Methods, Equation 10). Intuitively, the difference in the mean response to matches and distractors maps onto the amount of diagonal structure in a neuron's response matrix (shown). The linearly separable information values for the neurons in Fig 3a are labeled (I_L). **b)** The pairwise linear-nonlinear model (LN model) we fit to describe the transformation from IT to PRH, shown for two toy IT neurons. To create the LN model, pairs of IT neurons were combined via two sets of orthogonal linear weights, followed by a nonlinearity to create two model PRH neurons. **c)** To compare performance of the actual and model populations, Poisson trial-by-trial variability was regenerated for the actual IT and PRH populations from the mean firing rate responses across trials (the response matrix) for each IT and PRH neuron. Shown are ideal observer (left) and linear classification (right) performance of the following populations: IT (gray), PRH (black), the nonlinear (N) model

(gray dot-dashed), and the linear-nonlinear (LN) model (black dashed), with the same conventions described in Figure 4b.

Untangling can be attributed to a mechanism that combines IT neurons with asymmetric tuning correlations

To understand how the pairwise LN model untangles information, it is useful to first conceptualize how a nonlinearity can act to extract linearly separable information from a neuron's matrix. As described in Fig 6a, a nonlinearity can be effective in situations when the variance (i.e. the "spread") across one set of conditions (i.e. the matches) is higher than the other set (i.e. the distractors). In such scenarios, the nonlinearity can change a subset of responses within the high variance set and thus increase the difference between the mean response to matches versus distractors (which translates into an increase the amount of linearly separable target match information; Fig 5a). As a technical point, it is worth noting that variance differences don't always translate into the extraction of linearly separable information: initial differences in the average responses to matches versus distractors (i.e. linearly separable information that already exists in a neuron's matrix) can lessen the impact that the variance differences between matches and distractors will have when the nonlinearity is applied (Fig 6b; see Methods, Equation 20). In sum, nonlinearities are effective at extracting linearly separable information from a neuron's matrix when differences between the variance across the matches and the variance across distractors exist, under suitable conditions.

Our results (Fig 5c) suggest that pairing plays an important role in producing linearly separable information (as compared to applying a nonlinearity without pairing). How does pairing make a nonlinearity more effective? As described in Fig 6c, pairing is effective when two neurons have asymmetric tuning correlations for matches and distractors (e.g. a positive correlation, or similar tuning, for matches and a negative correlation, or the opposite tuning, for distractors). When two such neurons are combined, these tuning correlation asymmetries translate into variance differences between matches and distractors, and thus a scenario in which a nonlinearity can extract linearly separable information. For example, when two IT neurons are combined with positive weights, the positive tuning correlation between matches can preserve the variance across matches whereas the tuning anti-correlation between distractors can "squash" the variance across distractors, thus creating a variance difference that a nonlinearity can act upon (Fig 6c, top). Viewing the same computations from a slightly different perspective, we can envision the responses of these neurons in a population space similar to that depicted in Fig 4a (i.e. a population of size 2) where the representation of target matches and distractors is initially nonlinearly separable or "tangled" (Fig 6c, bottom left). Pairing the neurons via linear weights amounts to a rotation of the responses in the population space (Fig 6c, bottom center). If the two neurons have asymmetric correlations for matches and distractors, rotating this space will result in variance differences for matches and distractors when the population responses are projected along the marginals. Once the nonlinearity is applied, the new population space will be partially separable by a linear decision boundary or equivalently, the representation will be partially "untangled" (Fig 6c, bottom right).

We have formalized the intuitions presented in Fig 6c into a quantitative prediction of the amount of linearly separable information that can be gained by pairing any two IT neurons via an LN model of the form we fit to our data; our prediction relies on the degree of asymmetry in the neurons' match and distractor tuning correlations (see Methods, Equations 23, 25). Empirically we find that this prediction provides a good account of the linearly separable information extracted by our LN model of the transformation from IT to PRH (correlation of the actual and predicted information gains for each pair: r=0.84), confirming that the asymmetric tuning correlation mechanism is a good description of how the pairwise LN model "untangles" information.

This description of untangling via asymmetric tuning correlations reveals that for any given IT neuron, its best possible pair is one that has a perfect tuning correlation for one set (i.e. matches) and a perfect tuning anti-correlation for the other set (i.e. distractors). However, we note that any asymmetry in the tuning correlations for matches and distractors will translate into an increase in linear separable information, under appropriate conditions (see Fig 6b and Methods; Equation 20). To illustrate this, Fig 7a depicts the increase of linearly separable information as a function of the strength and sign of match and distractor tuning correlation pairings; note that the largest increases are found in the second and fourth quadrants of the plot (which correspond to opposite sign tuning correlations for matches and distractors), but increases of linearly separable information extend into the first and third quadrants (which correspond to same-sign tuning correlations) as well. How likely was it to encounter each type of pairing in our recorded IT population? Fig 7b shows a two-dimensional histogram of the correlations between all possible pairs of IT neurons. From this plot, one can see that match and distractor correlations are themselves correlated, and thus the maximally

effective pairings (the second and fourth quadrants) are rare whereas the maximally ineffective pairings (the first and third quadrants) are common. However, the LN model was able to capitalize on the modest tuning correlation asymmetries that did exist in IT (see the expected 2-D histogram of the LN model pairings in Fig 7c and the actual pairing in Fig 7d).

What produces these modest tuning correlation asymmetries in IT? Simple toy models (Fig 7e) and pseudosimulations (Fig 7f) reveal that asymmetric tuning correlations result when response matrices are mixtures of visual and target information. Notably, the existence of these asymmetries does not depend on the specific rules by which visual and target information are combined (i.e. congruently or incongruently).

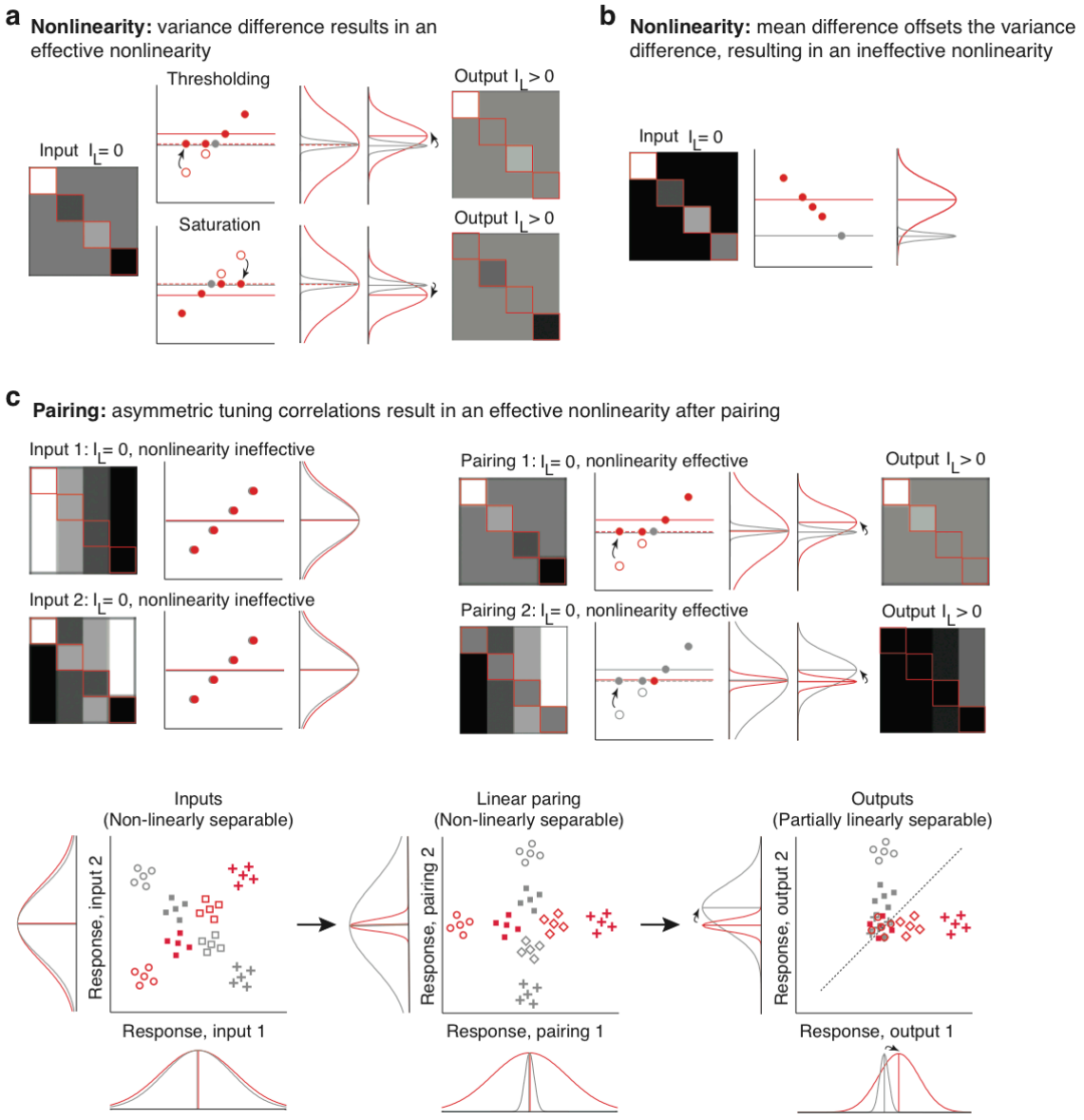

Figure 2-6. *Toy model description of the mechanisms underlying untangling.* **a)** *Shown from left-to-right*: 1) The response matrix for a toy model neuron with the match conditions outlined in red for visual effect. 2) The nonlinearity. Shown are plots of the responses of the original neuron (open circles) and the responses after the nonlinearity (solid circles) and the average firing rate before (dashed) and after (solid) the nonlinearity with the responses to match and distractor conditions shown in red and gray, respectively. 3) Gaussian distributions with the same means and variances for matches (red) and distractors (gray) as the input neuron; 4) Gaussian distributions with

the same means and variances for matches and distractors as the output neurons; 5) the response matrices for the output neurons. Shown is a toy model neuron that has the same average response to matches and distractors, and thus no linearly separable information (I_L=0). *Top,* However, because the lowest responses in the matrix are matches, a threshold nonlinearity can set these to a higher value, thus producing an increase in the overall mean match response such that it is now higher than the average distractor response. Because linearly separable information depends on the difference between these means, this translates directly into an increase in linearly separable information in the output (I_L>0). *Bottom,* Similarly, because the highest values of the response matrix are for matches, a saturating nonlinearity can decrease these responses, thus producing an decrease in the overall mean match response such that it is now lower than the average distractor response. In sum, the nonlinearity is effective because the variance is larger for the match as compared to the distractor condition. **b)** A toy model neuron for which the difference in the variances for matches and distractors is large, but because the means across matches and distractors are already offset, a nonlinearity is ineffective at extracting additional linearly separable information. As described in the methods (Equations 19-20), the amount of linearly separable information that can be extracted via a nonlinearity can be calculated based on the differences in the variances assuming no mean offset, followed by a correction that accounts for any mean offset that does exist. **c)** *Top left,* Two toy model neurons that have the same mean response to matches and distractors (hence no linearly separable information) and the same variance in their responses to matches and distractors (hence a nonlinearity applied to either of them would produce no increase in linearly separable information). *Top right,* Combining these neurons with positive weights maintains the variance in the responses to matches but decreases the variance in the responses to distractors, thus resulting in a scenario in which a nonlinearity is now effective. Similarly, combining these input neurons to produce a second model neuron using orthogonal weights maintains the variance to distractors while decreasing the variance to matches, thus producing a scenario in which a nonlinearity is effective. *Bottom,* The effectiveness of pairing can be attributed to an asymmetry (i.e. a difference) in the neurons' tuning correlations for matches and distractors (Methods, Equations 23, 25). Shown are the responses of each neuron in the same format as Fig 4a to illustrate how this mechanism

results in untangling within a population of two neurons. *Left,* The two toy neurons produce a nonlinearly separable representation in which a linear decision boundary is incapable of separating matches from distractors. However, these two toy model neurons have perfect tuning correlations for matches (i.e. both neurons respond to object 1 > 2 > 3 > 4 when presented as matches) and perfect tuning anti-correlations for distractors (i.e. neuron 1 responds to object 1 > 2 > 3 > 4 and neuron 2 responds to object 4 > 3 > 2 > 1 when presented as distractors). *Middle,* Pairing via linear weights produces a rotation within the two-dimensional space defined by the output of each member of the pair (see Methods). When the resulting responses are projected along the marginals (i.e. the responses of pairing 1 and pairing 2), this translates into variance differences for matches and distractors. Note that a linear decision boundary remains incapable of separating matches and distractors in this linearly transformed space. *Left,* Applying a nonlinearity to the linearly paired responses results in a representation in which a linear decision boundary is partially capable at distinguishing matches and distractors.

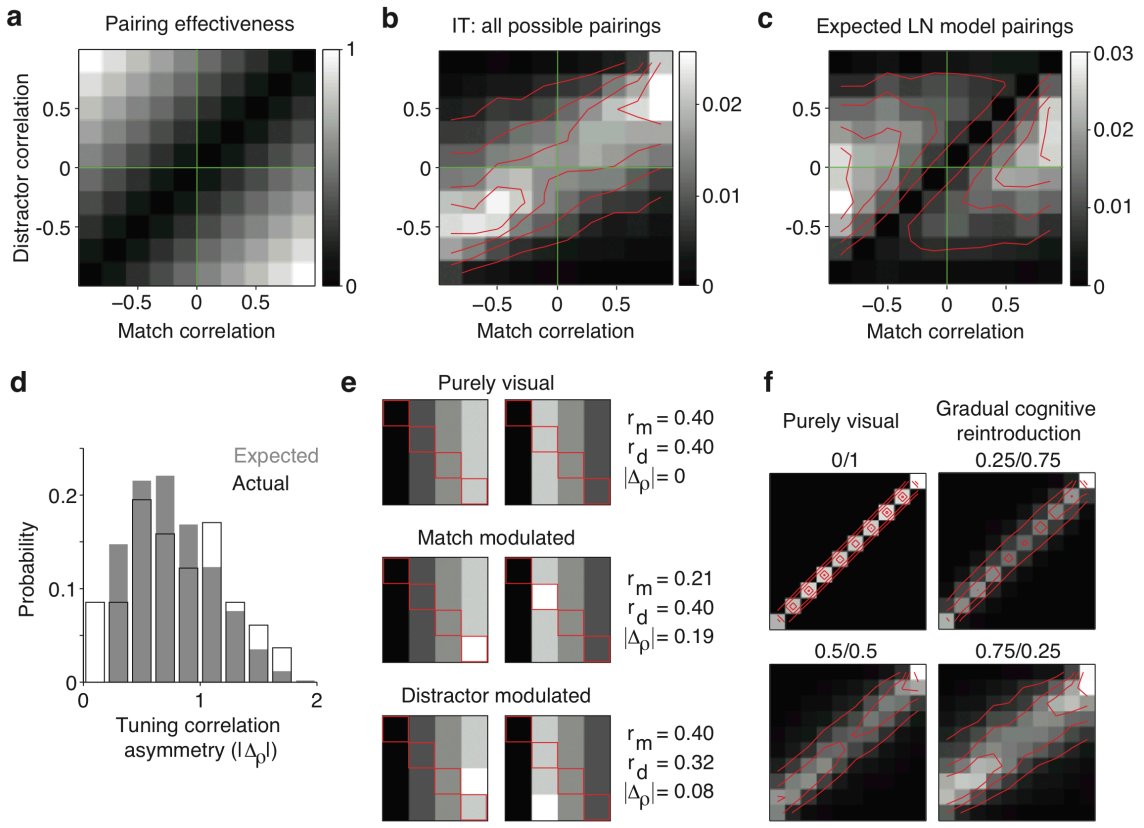

Figure 2-7. *Untangling largely relies on modest tuning correlation asymmetries.* **a)** The effectiveness of different possible match and distractor correlation pairings, computed as the increase of linearly separable information relative to the idealized case of perfect correlations for one set and perfect anti-correlations for the other, and plotted as a function of the correlation between matches and between distractors for a pair. **b)** Proportions of different match and distractor correlations computed for all possible pairs of IT neurons. **c)** Expected proportions of match and distractor correlation pairings, computed by weighting the distribution depicted in subpanel b by the distribution in subpanel a. **d)** Comparison of tuning correlation asymmetry (i.e. the absolute value of the difference between match correlations and distractor correlations) histograms for pairs chosen by the LN model (black) and the expected pairings based on the prediction in subpanel c (gray). The distributions are not statistically distinguishable as assessed by a comparison of their means (mean expected = 0.10; mean actual = 0.09; p=0.41) or a KS-test comparison of their cumulative probabilities (p=0.91); **e)** To illustrate the role that target modulations play in producing asymmetric tuning correlations for matches and

distractors, shown are response matrices for three pairs of toy model neurons. The match conditions are outlined in red for visual effect. *Top,* Two neurons with purely visual responses. The two neurons have the same responses to objects 1 and 3 and flipped responses for objects 2 and 4. Because visual matrices result in purely vertical matrix structure, this translates into the same correlation between matches (r_m, computed from the red entries of the matrix) and distractors ($r_{d,}$ computed from the other entries), and thus no tuning correlation asymmetry ($|\Delta_\rho|$). *Center,* Two neurons with tuning similar to those above, but each with match enhancement for their preferred object. This has the effect of decorrelating tuning for the matches, thus producing a tuning correlation asymmetry. *Bottom,* Similarly, if the two visual neurons (top) are enhanced for one distractor condition for their preferred object, this results in a tuning decorrelation for distractors, thus producing a tuning correlation asymmetry. **f)** To illustrate the role that "mixtures" of visual and cognitive signals play in shaping the histogram displayed in subpanel b, a pseudosimulation was performed in which a purely visual version of each neuron was created by assigning the responses of the neuron to each visual object (i.e. each column) to equal the average response to that object across all targets, thus producing a matrix with only vertical structure. Each IT neuron's matrix was then computed as a weighted sum of its actual matrix and the visual version of that matrix in the ratios depicted above each histogram plot (actual/visual). Compare also with subpanel b, which corresponds to a weight of 1/0. In subpanels b and c and f, red lines correspond to contours of constant proportionality at 0.2, 0.4, 0.6 and 0.8 the peak of the matrix.

Discussion

Finding specific visual targets requires the combination of visually selective and target selective signals. The ability to flexibly switch between different targets imposes the computational constraint that this combination must be followed by a refinement

process to signal whether a target is present in a currently-viewed scene (Fig 1). While the locus of the combination of visual and target-specific signals is thought to reside with mid-to-higher stages of the ventral visual pathway (Haenny, Maunsell et al. 1988, Maunsell, Sclar et al. 1991, Eskandar, Richmond et al. 1992, Gibson and Maunsell 1997, Liu and Richmond 2000, Chelazzi, Miller et al. 2001, Bichot, Rossi et al. 2005), the rules by which the brain combines and refines this information are not well understood. Our results build on earlier studies to: 1) provide a computational understanding that the rules of combination of visual and target-specific signals produce a nonlinearly separable or "tangled" representation of target matches that is then "untangled" in a higher brain area; and 2) provide a neural mechanism that can account for the untangling or refinement process. Notably, our results are not predictable from earlier reports. Specifically, a series of groundbreaking studies reported signals that differentiate target matches from distractors not only in PRH (Chelazzi, Miller et al. 1993, Miller and Desimone 1994), but also in V4 (Maunsell, Sclar et al. 1991) and IT (Eskandar, Richmond et al. 1992). Thus it has been difficult to discern the degree to which the target match signals present in PRH were directly inherited from IT (due to combinations of visual and target-specific information within or before IT that directly produced this type of signal), as compared to combinations of visual and target-specific information within or before IT that were then refined in PRH; our results provide evidence for the later hypothesis. Other more direct comparisons between IT and PRH have focused on aspects that are distinct from the task solution (e.g. responses during the cue or the delay period; Liu and Richmond 2000, Naya, Yoshida et al. 2003). Additionally, while untangling phenomena have been documented for perceptual tasks and in other brain areas (Hung, Kreiman et al. 2005, DiCarlo and Cox 2007, Rust and DiCarlo 2010), our

study is the first to propose a neural mechanism that can quantitatively account for this type of processing.

While not definitive, a number of lines of evidence support a model in which PRH untangles information arriving directly via inputs from IT. First, anatomical evidence suggests that the primary input to PRH is in fact IT (Suzuki and Amaral 1994). Second, our results demonstrate that nearly all the information for this task found in PRH is also contained in IT, suggesting that PRH need not get its input from other sources. Finally, our results demonstrate that a simple linear-nonlinear model can account for the transformation. Notably, the neural mechanism we have proposed constitutes a "functional model" of neural computation and this mathematical description would hold whether these computations are implemented within PRH or in another structure that receives input from IT and then passes this information to PRH.

Our results describe a mechanism by which information is reformatted to find specific targets - by combining neurons with asymmetric tuning correlations - and we anticipate that this mechanism can be generalized to account for the neural computations underlying other tasks as well. The "untangling" description has been most extensively applied to the computational challenge of constructing a neural signal that can identify an object invariant of its specific context (i.e. its retinal position or background; DiCarlo and Cox 2007), but the general concept applies to any situation in which the brain needs to create a representation from which a parameter can be easily extracted from a population response (e.g. determining the direction of a moving stimulus invariant of the stimulus pattern, Movshon, Adelson et al. 1985, or grouping stimuli into categories, Meyers, Freedman et al. 2008). However, describing computational mechanisms in

higher brain areas has proven to be a formidable challenge, due to the difficulties involved in constraining computational models in brain areas that lie far from the sensory inputs (DiCarlo, Zoccolan et al. 2012). We arrived at a mechanistic description of untangling for this particular problem (i.e. finding sought visual targets) by applying the type of linear-nonlinear model that has been vastly successful at describing computation at early stages of sensory processing (e.g. Carandini, Demb et al. 2005), using extensions that allow this type of model to be fit to higher brain areas (Rust, Mante et al. 2006). We do, however, caution that a determination of its applicability to other untangling problems requires a quantitative test.

Our results demonstrate that information is formatted in a manner more accessible to a simple (i.e. a linear) readout in PRH as compared to IT. For what purpose? While not definitive, consistent with a role in this circuit that transforms the stimulus into a behavioral response, we found that the PRH population representation of the task solution peaked well before the monkeys' reaction times (Fig 4e) and that the PRH representation differed on correct and error trials (Fig 4f). Other functional characterizations of PRH also support its role in contributing to the circuitry involved in computing reward or a decision during delayed-match-to-sample, target-switching tasks similar to the one we used in our experiment (Chelazzi, Miller et al. 1993, Miller and Desimone 1994, Liu and Richmond 2000). At the same time, our results and those of other studies suggest that PRH represents an intermediate, rather than the ultimate, stage of reward or decision computations. PRH projects to prefrontal cortex (PFC, Lavenex, Suzuki et al. 2002) and PFC neurons have been reported to convey more target match information than neurons in PRH (Miller, Erickson et al. 1996). Thus PRH reward-related signals may form an intermediate representation that is further refined in

PFC and used to guide behavior (as suggested by the model of Engel and Wang 2011). In agreement with these notions, we find that when we apply the pairwise LN model described in Fig 5 to our recorded PRH population (consistent with additional computations at stages beyond PRH), the resulting population performs significantly better than our recorded PRH (not shown). Our PRH population is thus likely to reflect one stage of the untangling process, and additional untangling computations are likely to occur at even later stages; future studies will be required to reveal whether this is in fact the case.

Supplementary Figures

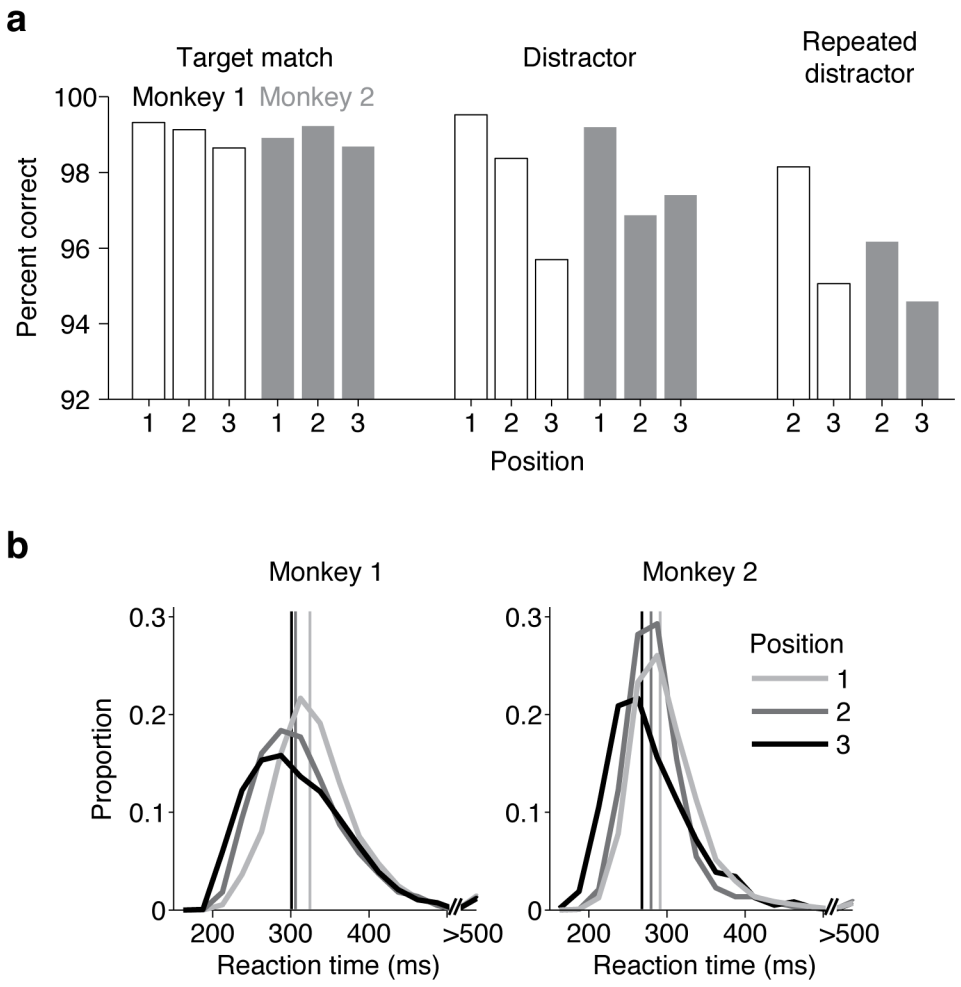

Figure 2-8. *Analysis of task performance.* **a)** Percent correct as a function of trial position for target matches, distractors, and repeated distractors. Note that these percentages are calculated per position, in contrast to the per trial calculations reported in the Results. Also note that most errors were "false alarms" (incorrect responses to a distractor) as compared to "misses" (incorrect responses to a target match). **b)** Distributions of reaction times as a function of position, calculated relative to stimulus onset. Vertical lines depict distribution means.

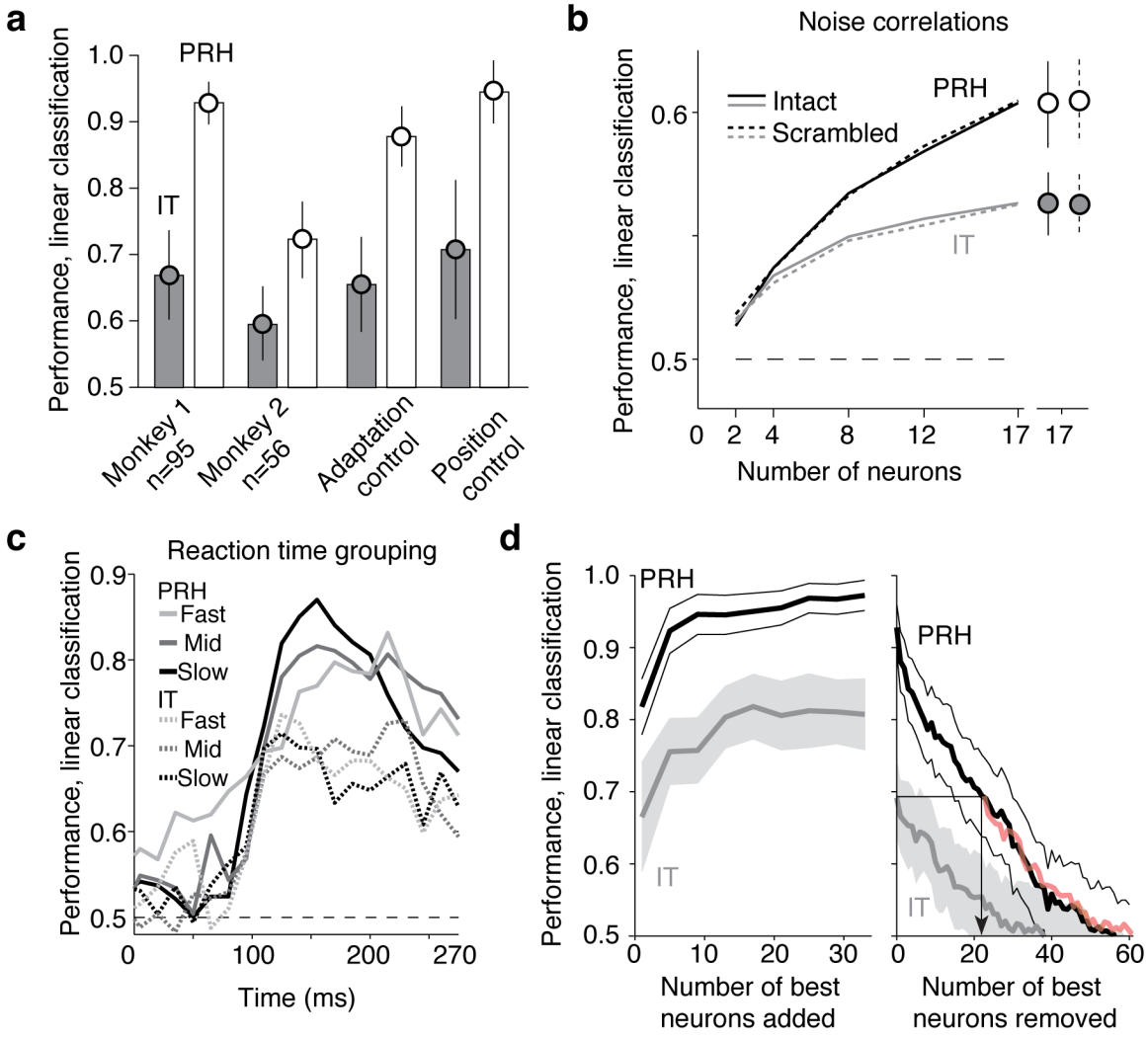

Figure 2-9. *Population performance controls.* **a)** Linear classification performance for a number of control conditions, including the IT (gray) and PRH (white) populations plotted for monkey 1 and monkey 2 individually. "Adaptation control": because the target match image is also presented earlier in the trial as the cue (Fig. 2), we determined whether repeated stimulus effects could account for discriminability for target matches and distractors by computing linear classifier performance for target matches and the second presentation of distractors that were presented twice in a trial (Fig. 2, "repeated distractor"). Performance remained higher in PRH as compared to IT, suggesting that

adaptation effects cannot account for the performance difference between these two areas. "Position control": a number of factors change as a function of the position a condition appears in a trial, including reward expectation (i.e. the probability that the next stimulus is a target match increases as a function of position). To determine whether position effects could account for the performance differences between PRH and IT, we computed linear classifier performance for discriminating target matches and distractors that were both presented as the first test stimulus after the cue. Note that due to a more limited number of trials available for the selection, spikes were counted in a shorter window (80 - 215 as compared to 80 - 270 ms) for the position-matched expectation control as compared to the other analyses**.** Performance remained higher in PRH as compared to IT, suggesting that position effects cannot account for the difference between these two areas. **b)** Comparison of linear classification performance for simultaneously recorded populations of size 17 (see Methods), when noise correlations (i.e. trial-by-trial variability) were left intact (solid) or scrambled (dotted). Shown is the average performance across 14 recording sessions in IT and 25 sessions in PRH. Note the similarity in performance for each population when noise correlations were kept intact and when they were scrambled **c)** Evolution of linear classification performance over time, as in Fig. 4a, but calculated with target match trials grouped by the monkeys' reaction times. To perform this analysis, we selected the target match trials as the four trials that corresponded to the fastest, middle, or slowest reaction times from each block (mean reaction times PRH fast = 261, middle = 305, slow = 370 ms; IT fast = 251, middle = 292, slow = 359 ms as compared to means of PRH = 334 and IT = 325 ms for the data presented in Fig. 4a). Note that for all three groupings, classification performance peaks in PRH before the monkeys' reaction times and that PRH

performance remains higher than IT performance. **d)** To relate population performance and single neuron responses, we performed analyses in which we systematically added (left) or removed (right) the N best neurons from the IT and PRH populations (where "best" was determined by the I_L metric plotted in Fig. 4c). *Left,* The arrow indicates that PRH performance was reduced to the values found in the full IT population when the best ~23 neurons were removed; the red trace indicates the IT performance plot shifted by 23 neurons. The standard error bars in panels a, b and d were computed with the conventions described for Fig. 3b

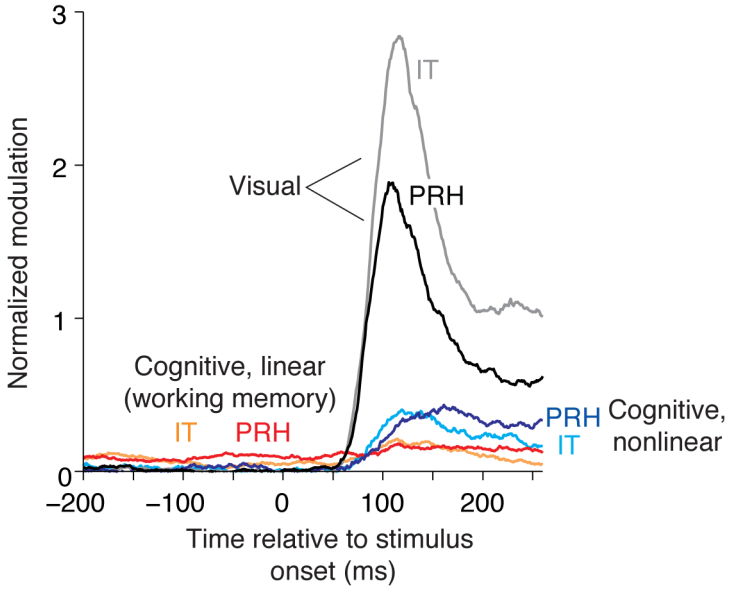

Figure 2-10. *Decomposition of cognitive information into its linear and nonlinear components.* Average magnitudes of visual (IT: gray, PRH: black), linear cognitive (IT: orange, PRH: red), and nonlinear cognitive (IT: cyan, PRH: blue) normalized modulation plotted as a function of time relative to stimulus onset (see also Fig. 5d,e). Normalized modulation was quantified as the bias-corrected ratio between signal variance and noise

variance (see Methods, Equation 4), and provided a noise-corrected measure of the amount of neural response variability that could be attributed to: “visual” - changing the identity of the visual stimulus; “linear cognitive” - changing the identity of the sought target (i.e. “working memory”); and “nonlinear cognitive” - nonlinear interactions between changing the visual stimulus and the sought target.

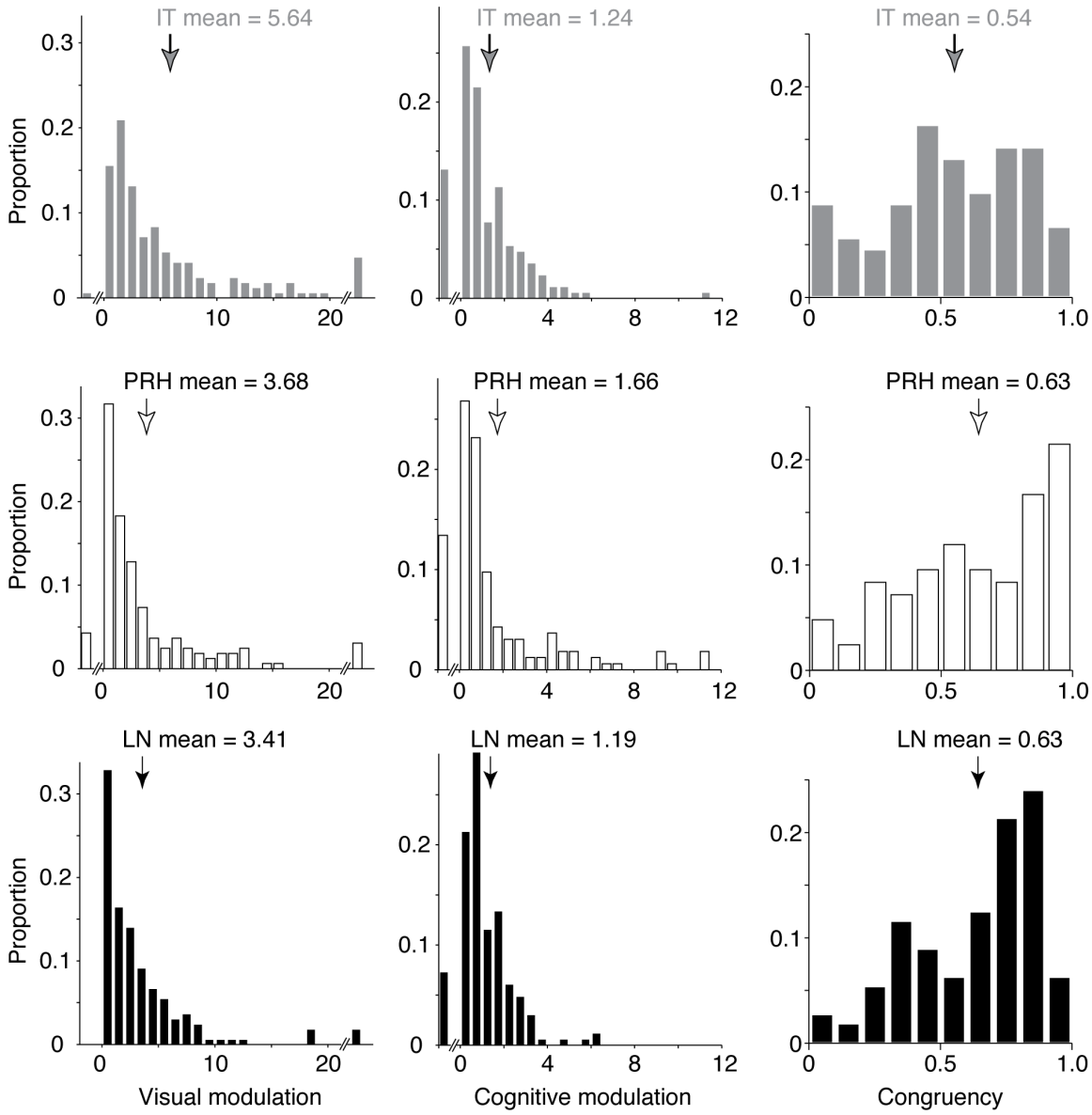

Figure 2-11. *The LN model (of PRH) accurately predicts differences between IT and PRH.* Histograms depict the distributions of: visual information (first column), cognitive information (second column) and congruency (third column) for: IT (first row, gray), PRH (second row, white) and the LN model (third row, black; Fig 6) populations. Visual and cognitive information were computed using a bias-corrected ANOVA analysis (see Methods, Equation 4). Congruency was designed to measure the degree of "alignment"

of visual and target modulation image preferences, and to range from 0, indicating complete misalignment, to 1, indicating perfect alignment (see Methods, Equations 5-8). Visual information was found to be significantly higher than in IT as compared to PRH ($p=0.0016$), and this difference was replicated by the LN model population ($p=0.002$). Cognitive information was not significantly different between IT and PRH ($p=0.056$) and this result was also replicated by the LN model population ($p=0.41$). Finally, congruency was found to be significantly higher in PRH as compared to IT ($p=0.006$), and this difference was replicated in the LN model population ($p=0.006$). The existence of incongruent neurons in IT could not be explained by neurons with poor single-unit isolation (Pearson correlation of the signal-to-noise ratio (SNR) measure of isolation and congruency: IT $r=0.04$; PRH $r=-0.001$) and we note that they have been documented by others as well [8, 20]. Neurons with values of visual information larger than 20 are included in the last bin of each histogram in the first column (proportions = 0.05 in IT, 0.03 in PRH, 0.01 in the LN model). The first (broken) bin of each histogram in the first two columns includes neurons with negligible amounts of visual or cognitive information (i.e. bias-corrected values lower than 0; proportions for visual information = 0.006 in IT, 0.04 in PRH and 0 in the LN model; proportions for cognitive information = 0.13 in IT, 0.13 in PRH and 0.07 in the LN model).

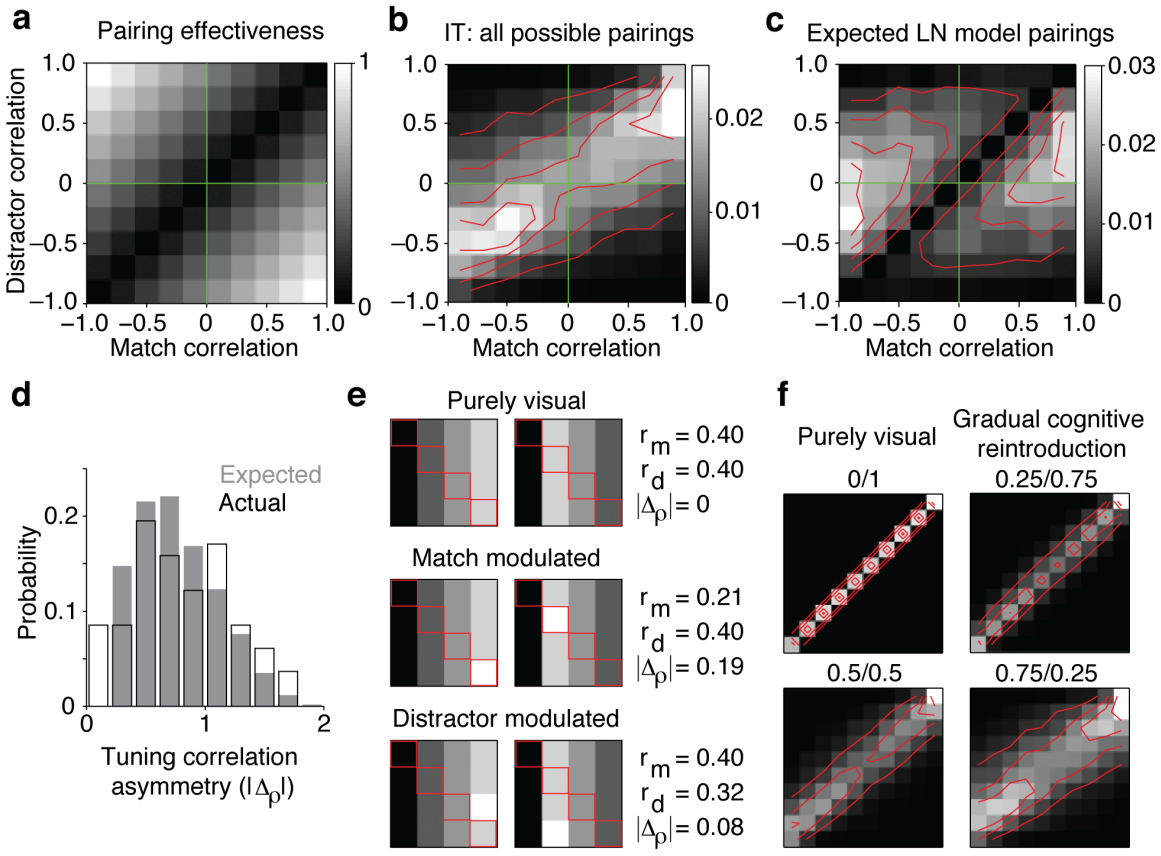

Figure 2-12. *Untangling largely relies on modest tuning correlation asymmetries.* **a)** To assess the types of tuning correlations that our pairwise LN model relied upon to untangle information (Figs. 6, 7), we began by determining the effectiveness of different hypothetical match and distractor correlation pairings. Effectiveness was computed as the increase of linearly separable information relative to the idealized case of perfect correlations for one set and perfect anti-correlations for the other, and is plotted as a function of the correlation between matches and between distractors for a pair. Note that the largest increases are found in the second and fourth quadrants of the plot (which correspond to opposite sign tuning correlations for matches and distractors), but increases of linearly separable information extend into the first and third quadrants (which correspond to same-sign tuning correlations) as well. **b)** Next we computed the

proportions of different match and distractor correlations for all possible pairs of our recorded IT neurons. In the recorded IT population, we found that match and distractor correlations were themselves correlated, and thus the maximally effective pairings (the second and fourth quadrants) were rare whereas the maximally ineffective pairings (the first and third quadrants) were common. **c)** Finally, we computed the pairings that were expected to be selected by the pairwise LN model (the maximally effective pairings that actually existed), by weighting the distribution depicted in subpanel b by the distribution in subpanel a. This analysis predicted that the LN model would rely on pairs of neurons with modest tuning correlation asymmetries. **d)** Comparison of tuning correlation asymmetry (i.e. the absolute value of the difference between match correlations and distractor correlations) histograms for pairs chosen by the LN model (black) and the expected pairings based on the prediction in subpanel c (gray). The distributions are not statistically distinguishable as assessed by a comparison of their means (mean expected = 0.10; mean actual = 0.09; p=0.41) or a KS-test comparison of their cumulative probabilities (p=0.91); **e)** To illustrate that modest tuning correlation asymmetries are ubiquitously present in populations of neurons that reflect mixtures of visual and target signals, shown are response matrices for three pairs of idealized neurons. The match conditions are outlined in red for visual effect. *Top,* Two neurons with purely visual responses. The two neurons have the same responses to objects 1 and 3 and flipped responses for objects 2 and 4. Because visual modulation produces purely vertical matrix structure, this translates into the same correlation between matches (r_m, computed from the red entries of the matrix) and distractors (r_d, computed from the other entries), and thus no tuning correlation asymmetry ($|\Delta_\rho| = 0$). *Center,* Two neurons with tuning similar to those above, but each with match enhancement for their preferred

object. This “congruent” target modulation has the effect of decorrelating tuning for the matches, thus producing a tuning correlation asymmetry. *Bottom,* Similarly, if the two visual neurons (top) are enhanced for one distractor condition for their preferred object, this results in a tuning decorrelation for distractors, thus producing a tuning correlation asymmetry. **f)** To illustrate the role that “mixtures” of visual and cognitive signals play in shaping the histogram displayed in subpanel b, a pseudosimulation was performed in which a purely visual version of each neuron was created by assigning the responses of the neuron to each visual object (i.e. each column) to equal the average response to that object across all targets, thus producing a matrix with only vertical structure. Each IT neuron’s matrix was then computed as a weighted sum of its actual matrix and the visual version of that matrix in the ratios depicted above each histogram plot (actual/visual). Note that a population of purely visual neurons (*top left*, weight 0/1) lacked asymmetric tuning correlations (i.e. all entries fall along the diagonal). However, gradual reintroduction of the actual target modulations resulted in a gradual reintroduction of tuning correlation asymmetries. Compare also with subpanel b, which corresponds to a weight of 1/0. In subpanels b and c and f, red lines correspond to contours of constant proportionality at 0.2, 0.4, 0.6 and 0.8 the peak of each matrix.

CHAPTER 3: Dynamic Target Match Signals in Perirhinal Cortex Can Be Explained by Instantaneous Computations That Act on Dynamic Input from Inferotemporal Cortex

Marino Pagan and Nicole C. Rust (2014). Journal of Neuroscience **34**(33): 11067-11084

Abstract

Finding sought objects requires the brain to combine visual and target signals to determine when a target is in view. To investigate how the brain implements these computations, we recorded neural responses in inferotemporal cortex (IT) and perirhinal cortex (PRH) as macaque monkeys performed a delayed match-to-sample target search task. Our data suggest that visual and target signals were combined within or before IT in the ventral visual pathway and then passed onto PRH, where they were reformatted into a more explicit target match signal over 10-15 ms. Accounting for these dynamics in PRH did not require proposing dynamic computations within PRH itself, but rather could be attributed to instantaneous PRH computations performed upon an input representation from IT that changed with time. We found that the dynamics of the IT representation arose from two commonly observed features: individual IT neurons whose response preferences were not simply rescaled with time and variable response latencies across the population. Our results demonstrate that these types of time varying responses have important consequences for downstream computation, and suggest that dynamic representations can arise within a feed-forward framework as a consequence of instantaneous computations performed upon time-varying inputs.

Introduction

Finding sought objects and switching between targets requires the flexible combination of visual information about the content of a currently viewed scene with working memory information about the identity of a sought target. These signals are thought to be combined within mid-to-higher stages of the ventral visual pathway (i.e. within V4 and IT; Fig 1), where the responses of neurons are modulated by changing both the identity of a visual stimulus as well as changing the identity of a sought target (Haenny, Maunsell et al. 1988, Maunsell, Sclar et al. 1991, Eskandar, Richmond et al. 1992, Gibson and Maunsell 1997, Liu and Richmond 2000, Chelazzi, Miller et al. 2001, Bichot, Rossi et al. 2005). The resulting target-modulated visual signals are then thought to be transformed into a "target match" signal that explicitly reports whether a currently-viewed scene contains a target via nonlinear computations that are implemented within PRH (Chelazzi, Miller et al. 1993, Miller and Desimone 1994, Pagan, L.S. et al. 2013), and prefrontal cortex (Miller, Erickson et al. 1996).

The computations required to create a target match signal can be envisioned as nonlinear conjunctions or "and-like" computations between visual and working memory signals (i.e. I am looking at my car keys "and" I am looking for my car keys). Evolution in the responses of neurons during and-like computations has been reported not only during target search (Chelazzi, Miller et al. 1993, Chelazzi, Miller et al. 2001) but also for computations involved in motion processing (Pack and Born 2001, Smith, Majaj et al. 2005) and object recognition (Brincat and Connor 2006). Specifically, these studies have revealed a delay between the time signals arrive within a brain area and the time that conjunction information appears on the order of tens of milliseconds. These delays have

been attributed to the time required for recurrent circuits within the brain area performing the computation to execute it (Brincat and Connor 2006), possibly via a biased, competitive process (Chelazzi, Miller et al. 1993).

Here we report a similar phenomenon, but one in which delays in the emergence of conjunction information can be attributed to computations that are instantaneous and fixed but act on an input representation that changes over time. More specifically, we found that during a delayed match-to-sample target search task, visual and working memory signals were partially represented in PRH as separate signals that then evolved into an "and-like" target match signal over ~10-15 ms. These dynamics were not simply inherited from the IT inputs and surprisingly, they did not require proposing dynamic computation within PRH itself. Rather, our data were well accounted for by a description in which the "and-like" computations that produce target match signals in PRH were gradually implemented across at least two processing stages: one that combined visual and working memory signals within or before IT to form a nonlinearly separable and time-varying representation, followed by computations in PRH that instantaneously reformatted the input arriving from IT to produce a more explicit target match signal.

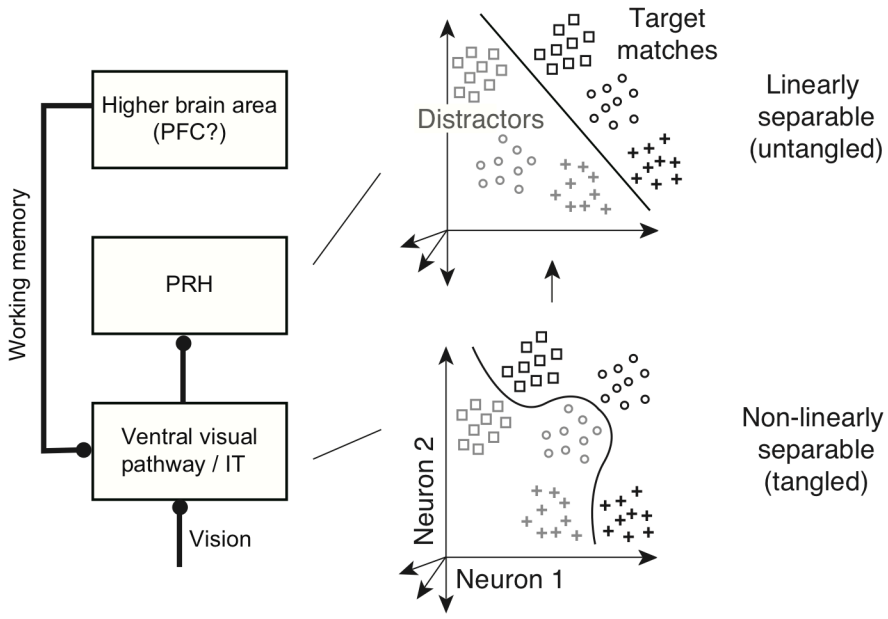

Figure 3-1. *"Untangling" target match signals. Left,* Previous results suggest that during visual target search, visual and working memory signals are combined within or before IT along the ventral visual pathway in a nonlinearly separable or "tangled" fashion, followed by computations in PRH that "untangle" target match information such that it is more accessible to a linear population read-out. *Right,* Each point depicts a hypothetical population response, consisting of a vector of the spike count responses to a single condition on a single trial. Clouds of points depict the predicted dispersion across repeated presentations of the same condition due to trial-by-trial variability. The different shapes depict the hypothetical responses to different images and the two shades (black, gray) depict the hypothetical responses to target matches and distractors, respectively. A target-switching task (such as the delayed match-to-sample task, Fig 2) requires discriminating the same objects presented as target matches and as distractors. In a "tangled" representation (bottom), a nonlinear decision boundary (corresponding to a nonlinear population read-out) is required to separate these two groups whereas an "untangled" representation (top) can be read-out with a linear decision boundary (corresponding to a linear population read-out). As reported by Pagan et al (2013), target match signals are more "tangled" in IT and more "untangled" in PRH.

Methods

The data reported here are the same data described by Pagan et al. (2013). The experimental procedures involved in collecting the data are described in detail in that report and are summarized here. Experiments were performed on two naive adult male rhesus macaque monkeys (*Macaca mulatta)* with implanted head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Monkeys performed a delayed match-to-sample task (Fig. 2a). Monkeys initiated each trial by fixating a small dot. After a 250 ms delay, an image indicating the target was presented, followed by a random number (0-3, uniformly distributed) of distractors, and then the target match. Each image was presented for 400 ms, followed by a 400 ms blank. Monkeys were required to maintain fixation throughout the distractors and make a saccade to a response dot located 7.5 degrees below fixation after 150 ms following target match onset but before the onset of the next stimulus to receive a reward. The same four images were used during all the experiments. Approximately 25% of trials included the repeated presentation of the same distractor with zero or one intervening distractors of a different identity. Behavioral performance was high (monkey 1=94%; monkey 2=92%). The same target remained fixed within short blocks of ~1.7 minutes that included an average of 9 correct trials. Within each block, 4 presentations of each condition (for a fixed target) were collected and all four target blocks were presented within a "metablock" in pseudorandom order before reshuffling. A

minimum of 5 metablocks in total (20 correct presentations for each experimental condition) were collected. The main components of this experimental design included 16 different conditions that could be envisioned as existing within a 4x4 matrix defined by each of the four images presented as a visual stimulus in the context of looking for each of the four images as a target (Fig 2b). This matrix includes four "target match" conditions, which fall along the diagonal of this matrix and twelve "distractor" conditions, which fall off the matrix diagonal.

Both IT and PRH were accessed via a single recording chamber in each animal. Chamber placement was guided by anatomical magnetic resonance images and later verified physiologically by the locations and depths of gray and white matter transitions. The region of IT recorded was located on both the ventral superior temporal sulcus (STS) and the ventral surface of the brain, over a 4 mm medial-lateral region located lateral to the anterior middle temporal sulcus (AMTS) that spanned 14-17 mm anterior to the ear canals (Liu and Richmond 2000, Rust and DiCarlo 2010). The region of PRH recorded was located medial to the AMTS and lateral to the rhinal sulcus and extended over a 3 mm medial-lateral region located 19-22 mm anterior to the ear canals (Liu and Richmond 2000). We recorded neural activity via a combination of glass-coated tungsten single electrodes (Alpha Omega, Inc.) and 16- and 24-channel U-probes with recording sites arranged linearly and separated by 150 micron spacing (Plexon Inc.). Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored via the OmniPlex Data Acquisition System (Plexon, Inc.). We performed all spike sorting manually offline using commercially available software (Plexon, Inc.).

Responses were only analyzed on correct trials. Target matches that were presented after the maximal number of distractors (n=3) occurred with 100% probability and were discarded from the analysis. The response of each neuron was measured as the spike count in time bins 25 ms wide and sampled at 1 ms intervals aligned to the onset of each visual image. In some of our analyses (described below), we assume that trial-by-trial response variability arose from a Poisson process, as we found this to be a good account of our data. For each neuron at each bin position (-50 to 250 ms relative to stimulus onset), we estimated the Fano factor by fitting the relationship between the mean and variance of spike counts for each of the 16 experimental conditions (Rust, Schultz et al. 2002). Grand mean Fano factor estimates averaged across all neurons and all windows (based on spike counts in 25 ms windows with shifts of 1 ms) was 1.01 in both IT and PRH. Similar to other reports (Churchland, Yu et al. 2010), we found a small but reliable decrease in Fano factor following stimulus onset (e.g. in IT average Fano factor dropped from a maximum of mean +/- sd of 1.06+/-0.15 at -50 ms to 0.94+/-0.13 at 112 ms).

Population performance

To measure the amount and format of information available in the IT and PRH populations to discriminate target matches and distractors, we performed two cross-validated classification analyses: a linear read-out and an ideal observer read-out (Pagan, L.S. et al. 2013). For both analyses, we considered the spike count responses of a population of N neurons (where N=164 in IT and PRH) to each condition as a

population "response vector" $\boldsymbol{x}$ with dimensionality equal to Nx1. Our experimental design resulted in 4 target match conditions and 12 distractor conditions; on each iteration we randomly selected 1 distractor condition from each image (for a total of 4 distractor conditions) to avoid artificial overestimations of classifier performance that could be produced by taking the prior distribution into account (e.g. scenarios in which the answer is more likely to be "distractor" than "target match"). The linear read-out (Fig 3a) amounted to finding the linear hyperplane that would best separate the population response vectors corresponding to all of the target match conditions from the response vectors corresponding to all of the distractor conditions and took the form:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b \quad (1)$$

where $\boldsymbol{w}$ is a Nx1 vector describing the linear weight applied to each neuron and b is a scalar value that offsets the hyperplane from the origin and acts as a threshold. The population classification of a test response vector was assigned to a target match when $f(\boldsymbol{x})$ exceeded zero and was classified as a distractor otherwise. The hyperplane and threshold for each classifier were determined by a support vector machine (SVM) procedure using the LIBSVM library (http://www.csie.ntu.edu.tw/~cjlin/libsvm/) with a linear kernel, the C-SVC algorithm, and cost (C) set to 0.1.

Our "ideal observer" read-out (Fig 3a) was designed to be "ideal" in the sense that its performance was limited by the amount of overlap in the trial-by-trial responses to target matches and distractors (i.e. it is optimal under the assumption of Poisson trial-by-trial variability) but not by the complexity of the decision boundary required to connect the multiple target match conditions and parse those from the multiple distractor conditions. We note that the "ideal observer" is not proposed as a neurally plausible

read-out, but rather as a method to estimate the maximum achievable performance using an arbitrarily complex read-out. To distinguish it from read-outs that impose a particular decision boundary (i.e. "linear") we refer to it as a measure of "total" information. Importantly, this ideal observer will perform well under a range of circumstances in which complete information for this task exists (e.g. at one extreme, a population of individual neurons that each convey large amounts of linearly separable target match information; and at the other extreme, a population that contains visual and working memory signals in separate subpopulations of neurons). Additionally, this ideal observer will fail under conditions in which target match information is incomplete (e.g. a population that contains purely "visual" or "working memory" neurons alone). To determine the ideal observer read-out, we computed the average spike count response r_{uc} of each neuron u to each condition c. The likelihood that a test response k arose from a particular condition for a neuron was computed as the Poisson probability density:

$$lik_{u,c}(k) = \frac{(r_{uc})^k \cdot e^{-r_{uc}}}{k!} \tag{2}$$

When applied to our model responses (see Methods, "Model Structure"), the Poisson probability density was extended to continuous responses by replacing the factorial with the Gamma function (note that this formula is equivalent to Equation 2 when k is an integer):

$$lik_{u,c}(k) = \frac{(r_{uc})^k \cdot e^{-r_{uc}}}{\Gamma(k+1)} \tag{3}$$

The likelihood that a test response vector $\boldsymbol{x}$ arose from each condition c for the population was computed as the product of the likelihoods for the individual neurons:

$$lik_c(\boldsymbol{x}) = \prod_u lik_{u,c}(x_u) \tag{4}$$

where x_u indicates the response of unit u on a single trial. Finally, we computed the likelihood that a test response vector arose from the category "target match" versus the category "distractor" as the mean of the likelihoods for target matches and distractors, respectively:

$$lik_{Match}(\boldsymbol{x}) = \frac{1}{4} \cdot \sum_{c \in Match} lik_c(\boldsymbol{x}) \quad ; \quad lik_{Distractor}(\boldsymbol{x}) = \frac{1}{4} \cdot \sum_{c \in Distractor} lik_c(\boldsymbol{x}) \tag{5}$$

The population classification was assigned to the category with the higher likelihood.

For both types of classifiers, we computed cross-validated performance by randomly assigning 50% of our data (10 repeats) to compute the representation ("training set") and testing with the remaining 50% of our data (10 repeats). To compute performance mean and standard error we performed a resampling procedure in which we randomly assigned repeats without replacement for training and testing. To combine the responses of neurons recorded in different sessions into a pseudopopulation, on each bootstrap iteration we shuffled the trial pairings between neurons to destroy any (artificial) trial-by-trial correlation structure. The read-out was trained separately for each timepoint, but across different timepoints for the same neuron, we always analyzed data from the same experimental trials. We performed 3000 resampling iterations for each timepoint. Estimates of the mean and standard error of performance at each timepoint were obtained by computing the mean and standard deviation across bootstrap iterations (e.g. Fig 3a).

To compute latency estimates for each type of classification (i.e. the latency for performance to reach a criterion; Fig 3b,c), we considered the performance values *p* for all timepoints *t* on one bootstrap iteration and we fit a 12th order polynomial to that data by minimizing mean square error:

$$p = \sum_{i=0}^{12} a_i t^i \tag{6}$$

(i.e. the function "polyfit" in Matlab). We used the resulting function to compute the latency values that corresponded to a range of criteria (i.e. the first timepoints that corresponded to performance values ranging from 0.55-0.875), although we could not estimate latencies on bootstrap iterations in which a criterion exceeded the maximum of the fitted function. We computed the latency mean and standard error for each criterion as the mean and standard deviation across these latency estimates. We computed the p-value for each criterion by considering pairs of latencies for the ideal observer and linear classifier and determining the fraction of those pairs for which the difference was flipped in sign relative to the actual difference between the means of the full dataset (i.e. the fraction of bootstrap iterations in which the ideal observer classification latency was larger than the linear classification latency; Efron and Tibshirani 1994). Additionally, we determined the degree to which smaller magnitude linear classifier performance could account for its longer latency relative to ideal observer performance by selecting the subset of bootstrap iterations on which ideal observer and linear classifier performance had the same distribution of magnitudes within a window of 135-140 ms, and we then calculated latencies on those magnitude-matched trials (Fig 3b-c). Specifically, we performed a histogram equalization in which we computed histograms of performance

averaged from 135-140 ms for both classifiers, and within each histogram bin, we randomly selected the same number of entries from each distribution. We then used the data from earlier timepoints on the same trials as these entries to calculate the mean and standard errors for latencies as described above.

By design, our ideal observer classifier (designed to measure "total" information) is capable of retrieving a more complex decision boundary than the linear classifier (i.e. because "linear" is a subset of "total"). This is reflected in the larger number of degrees of freedom available to the ideal observer. Specifically, the number of ideal observer degrees of freedom is equal to the number of neurons multiplied by the number of discriminated conditions (i.e. 164*8, given that 4 matches and a subset of 4 distractors where discriminated on each bootstrap iteration of the classification procedure as described above) whereas the number of linear classifier degrees of freedom is equal to the number of neurons (i.e. 164). Because we indirectly infer the timecourse of nonlinearly separable information by comparing ideal observer and linear classifier performances, we designed a control analysis to evaluate whether differences in the numbers of parameters led to a spurious interpretation of our results. To do this, we developed a new linear and nonlinear classifier designed to measure each of these quantities directly and with the same number of parameters. The approach we took is analogous to a polynomial expansion of the classifier read-out rule, where the first term corresponds to a linear classifier that captures differences between the mean population responses, and the second term correspond to a nonlinear (quadratic) classifier chosen to maximize the difference between the variances of the population responses. Specifically, this "linear" classifier operates by maximizing the distance between the mean response across all matches and the mean response across all distractors and

was computed as the difference between the population response vector averaged across all matches and the population response vector averaged across all distractors (for the training data). Next, a threshold was computed via a brute-force search as the value that maximizes the fraction of correct classifications of matches and distractors in the training data. In contrast, the nonlinear classifier operates by projecting the training data onto the axis that maximizes the difference between the variance in the response across all matches and the variance in the response across all distractors and this vector was computed from the eigendecomposition of the difference between the covariance matrices for matches and for distractors computed from the training data; this nonlinear classifier was taken as the eigenvector with the maximum absolute eigenvalue. After the population responses were projected along this axis, they were squared (which acts to convert these variance differences into mean differences) and a final threshold was computed as the value that maximized the fraction of correct classifications of matches and distractors in the training data. We note that the number of free parameters employed by both classifiers is the same and is equal to the number of neurons in the population (i.e. one weight for each neuron = 164). The two classifiers also have a similar structure, consisting of a dot product between the weight vector and the population response followed by thresholding, and the only difference between the two classifiers is a parameter-free squaring operation for the nonlinear classifier that is applied before thresholding. The same cross-validation procedure described above for the ideal observer and SVM linear classifier was used to compute mean and standard error of performance for these linear and nonlinear classifiers (Fig 3d) and as was the case for the ideal observer and SVM linear classifier, the parameters for these linear and nonlinear classifiers were optimized for each time bin.

Decomposition of single-neuron responses

We applied a method to decompose the response matrix for each neuron into modulations along a fixed set of intuitive, task-relevant components (Pagan and Rust 2014): visual stimulus identity ("visual"), target identity ("working memory"), whether each condition was a target match or a distractor ("diagonal"), and all other nonlinear combinations of visual and working memory modulations ("non-diagonal"; Fig 4a). We also use the term "cognitive" to indicate the combined working memory and non-diagonal signals. Our method bears some resemblance to a classic analysis of variance (ANOVA). However, a two-way ANOVA applied to our data would parse each response matrix into "visual", "working memory" and "nonlinear interaction" terms and for our task, differentiating among different types of nonlinear interaction terms (e.g. diagonal versus non-diagonal) is crucial. Our analysis is also similar to a principal components analysis (PCA), which recovers a set of orthonormal basis components that capture the response modulations of a population by assigning each successive component to account for as much of the remaining population response variance as possible. However, PCA components are not guaranteed to be intuitive whereas our method involves fixing the components to account for intuitive parameters and quantifies the magnitude of response modulation along each of them. To obtain the basis functions, we first defined a set of 16 linearly independent matrices whose entries differentiated between different conditions and we then applied the Gram-Schmidt process to impose that each matrix had unitary Euclidean norm and that all matrices were orthogonal. The resulting

orthonormal basis is shown in Figure 7b. It is worth noting that while the specific basis functions used to describe these modulation components are not unique (e.g. one could define another set of orthogonal vectors that would capture the visual modulations equally well), the linear subspaces captured by these specific subsets of components (e.g. the three visual components, Fig. 7b) are uniquely defined. This follows from the inherent two-dimensional "looking at" / "looking for" matrix structure this task (Fig 2b), in which the "visual" and "working memory" conditions are presented in all possible combinations and are thus independent from one another. In other words, the combined projection of a neuron's response vector onto the three visual components uniquely captures the amount of modulation that can be attributed to changes in the identity of the visual stimulus.

A neuron's response matrix $\boldsymbol{R}$ can be decomposed into a weighted sum of these components:

$$\boldsymbol{R} = \sum_{i=1}^{16} m_i \cdot \boldsymbol{b_i} \tag{7}$$

where $\boldsymbol{b_i}$ indicates the i-th component, and m_i indicates the weight (i.e. the amount of modulation) associated with the i-th component. Components of the same type (i.e. the three visual components in Fig. 7b) can then be grouped together to quantify the amount of each type of task-relevant modulation. More specifically, each type of modulation can be computed as the square root of the sum of the squared modulations along all relevant components:

$$M_{vis} = \sqrt{\sum_{i \in vis} m_i^2} \quad ; \quad M_{wm} = \sqrt{\sum_{i \in wm} m_i^2} \quad ; \quad M_{diag} = \left| m_{diag} \right| \quad ; \quad M_{non-diag} = \sqrt{\sum_{i \in non-diag} m_i^2}$$

(8)

where M_{vis} is the amount of visual modulation, M_{wm} is working memory modulation, M_{diag} diagonal modulation, and $M_{non-diag}$ non-diagonal modulation.

Bias correction of response components

When estimating the amount of modulation (or information) in a signal, noise and limited sample size are known to introduce a positive bias (e.g. Treves and Panzeri 1995). For example, consider an hypothetical neuron that responds with the same average firing rate response to each of a set of stimuli. Because neurons are noisy, if we were to estimate these mean rates based on a limited number of repeated presentations, we would get the erroneous impression that the neuron does in fact differentiate between the stimuli. To overcome this problem, we estimated this bias using a bootstrap procedure and corrected for it. By reversing Equation 7, the estimated squared modulation along each component i is given by:

$$m_i^2 = \left(\boldsymbol{R} \cdot \boldsymbol{b_i}^T\right)^2 = \left(\sum_{j=1}^{16} r_j \cdot b_{ij}\right)^2 \quad \textbf{(}9\text{)}$$

where r_j indicates the neuron's average response to the j-th condition, and b_{ij} indicates the j-th entry of the i-th basis component. To estimate the bias introduced by limited sampling, we applied a bootstrap procedure in which we first resampled with replacement 20 responses to each condition and we then recomputed the squared

modulation of these bootstrapped responses. The bias could be estimated by subtracting the modulation computed on the actual responses from the bootstrapped modulation (Efron and Tibshirani, 1994):

$$Bias_i = \hat{m}_i^2 - m_i^2 \tag{10}$$

where $\hat{m}_i^2$ indicates the squared modulation computed on the resampled responses. Bias was independently computed and subtracted from each type of modulation. Using procedures described in detail in (Pagan and Rust 2014), we having confirmed the validity of this bias correction procedure and its equivalence to other bias correction approaches for this spike count window size, numbers of trials, and specific data set.

Relationship between single-neuron responses and population performance

The population performance of a linear classifier for discriminating target matches from distractors depends on the total amount of diagonal modulation (i.e. the differences in the firing rate responses to target matches as compared to distractors). We define the total amount of diagonal modulation in a population $M_{diag,pop}$ as the square root of the sum of the squared diagonal modulation $M_{diag,n}$ for each neuron n:

$$M_{diag,pop} = \sqrt{\sum_n M_{diag,n}^2} \tag{11}$$

To transform this measure into an estimate of the performance of a linear classifier, $Perf_{SVM}$, we applied the following formula (Poor 1994, Averbeck and Lee 2006):

$$Perf_{SVM} = 1 - H\left(\frac{k_{eff} \cdot M_{diag,pop}}{2}\right) \quad (12)$$

where H is the complementary error function and k_{eff} is a classifier efficiency factor applied to account for the inability of the classifier to extract all the available information (for example, due to the limited amount of training data resulting in suboptimal choice of the parameters). This efficiency parameter is mathematically equivalent to the one introduced by Giesler and Albrecht (1997), although applied for a slightly different purpose in that case (to relate neural responses and behavior). We empirically estimated k_{eff} as 0.49 and we applied this same value for the estimation of both the linear classifier and the ideal observer (described below). In the main text, we use the term "linear classifier component" to refer to the quantity $k_{eff} \cdot M_{diag,pop}$.

The performance of an ideal observer for discriminating target matches from distractors can be approximated using the sum of the linear classifier component and a "nonlinear classifier component" $k_{eff} \cdot M_{NL-diag,pop}$ that reflects the amount of nonlinearly separable target match modulation contained in the combined visual and cognitive signals:

$$M_{NL-diag,pop} = \sqrt{\frac{M_{vis,pop}^2 \cdot M_{cog,pop}^2}{M_{vis,pop}^2 + M_{cog,pop}^2}} \quad (13)$$

where $M_{vis,pop}$ is computed from each neuron's visual modulation analogously to $M_{diag,pop}$, and $M_{cog,pop}$ measures the amount of cognitive modulation as the sum of working memory and non-diagonal modulation:

$$M_{vis,pop} = \sqrt{\sum_n M_{vis,n}^2} \quad ; \quad M_{cog,pop} = \sqrt{\sum_n \left(M_{wm,n}^2 + M_{non-diag,n}^2 \right)} \tag{14}$$

Finally, the performance of an ideal observer $Perf_{ID.OBS.}$ can be estimated as:

$$Perf_{ID.OBS.} = 1 - H\left(\frac{k_{eff} \cdot M_{diag,pop} + k_{eff} \cdot M_{NL-diag,pop}}{2} \right) \tag{15}$$

Comparison between estimated and actual performances for our recorded neurons are shown in Figs. 4c and 5c, the magnitudes of linear and nonlinear classifier components are shown in Fig. 4d, and the mapping function used to transform classifier components into estimated performances in Equations 12 and 15 is plotted in Fig. 4e.

Fitting an instantaneous feed-forward model of PRH to the IT responses

Our goal was to fit an instantaneous linear-nonlinear (LN) model to responses of IT neurons to determine whether this type of model could reproduce the dynamics observed in our recorded PRH population. To constrain the model, we assumed that the brain implements this transformation optimally and we thus determined the model parameters that maximized the total amount of diagonal modulation $M_{diag,pop}$ in our model PRH. We always performed this maximization at a single timeslice relative to

stimulus onset, and we explored different positions of that training window (e.g. 75 versus 135 ms, Fig 6b-c). Similar to the classification procedures described above, our model was designed to be cross-validated. On each iteration of the cross-validation procedure, 50% of the IT responses (10 repeats) were used to determine the LN model parameters, while the remaining 50% were passed through the instantaneous LN transformation to produce a set of "model PRH responses". The model PRH responses were then compared to the actual PRH responses by measuring the performances of the same linear classifier and ideal observer described above (Fig. 6). The cross-validation of the model and the classifier were integrated, so that the same repeats used to train the model parameters were also used to train the classifier parameters, while the "test repeats" (i.e. the model PRH responses) were used to determine the classifier performances.

Model structure

The responses of each model PRH neuron were created as an n-way linear combination of n IT responses, followed by a static nonlinearity, where n corresponds to the total number of neurons in our IT population (n=164). The linear transformation was applied to individual trials (i.e. to the spike counts obtained for each of the conditions in one randomly selected repeat of the response matrix $\boldsymbol{R_i}$ for each i-th IT cell) to produce a new matrix L :

$$L = \sum_i w_i \cdot \boldsymbol{R_i} \tag{16}$$

where w_i indicates the weight applied to the i-th IT neuron. The vectors of weights applied to create different PRH model neurons were constrained to be orthogonal and to have unitary norm:

$$\overline{W_i} \cdot \overline{W}_j = 0 \qquad ; \qquad \sum_i w_i^2 = 1 \tag{17}$$

where $\overline{W_i}$ and $\overline{W_j}$ are the vectors of weights for the i-th and the j-th model neurons. The matrix L resulting from Equation 16 was then passed through an instantaneous static nonlinearity to produce the model responses for the matrix on a single trial (described below). The trial-by-trial variability in the resulting model PRH thus arose from the trial-by-trial variability recorded in IT.

Fitting procedure

The fitting procedure was designed to determine the linear weights for each model PRH neuron with the goal of maximizing the overall diagonal modulation in the model population. Maximizing diagonal modulation required us to generate model neurons via linear combinations of IT responses that both preserved the diagonal modulation already present in the input as well as extracted the maximal new diagonal signal (once nonlinearities were applied). We achieved this by splitting these two types of signals into two different classes of model neurons and together, these two classes fully captured all the information available within the IT responses at the timepoint used to train the model (described in detail below). A third class of model neurons captured

all remaining information present at all other timepoints (described below). This approach, which involved splitting the total amount of information into separate (linear and nonlinear) terms, is analogous to the linear and nonlinear classifiers with matched numbers of parameters introduced above.

To determine the parameters for the model, we began by normalizing the responses of each IT neuron on individual trials by subtracting the grand mean across all conditions and dividing the result by the standard deviation across trials, pooled across all conditions. This normalization helped to ensure that the linear weights were assigned based on a measure that reflected both the magnitude of responses as well as trial-by-trial variability (i.e. d-prime) as opposed to raw spike count responses alone. These "normalized responses" were used to find the linear weights, and once determined, the weights were converted back to units of spike count before the nonlinearities were applied; we note that the "normalized responses" were used only to fit the model parameters whereas "un-normalized" spike counts were used to determine the cross-validated responses of the model itself.

The first model neuron was fit with the goal of preserving the diagonal signals contained in the recorded IT responses. This was achieved by choosing the linear weights for this neuron as the optimal linear discriminant between target match and distractor normalized responses (i.e. the vector of weights that connects the mean normalized response to target matches and the mean normalized response to distractors; Fig 6a, left). The weights were then unnormalized, and the result of the linear combination with these weights L (computed according to Equation 16) was centered

(by subtracting the mean μ), and exponentiated to produce the final response matrix LN_1:

$$LN_1 = \exp(L - \mu) \tag{18}$$

The monotonicity of the exponential function ensures that the rank-order of match and distractor responses is preserved, while at the same time making all responses positive.

The sets of linear weights for the second class of model PRH neurons were determined with the goal of maximizing the amount of diagonal modulation that could be extracted from the remaining IT population response space (i.e. after the axis defined by the weights of the first model neuron was removed, thus reducing the dimensionality by 1). The intuition behind the process used to extract diagonal information has been described in our previous report (Pagan et al., 2013), and is briefly summarized in Fig. 6a, center. The key step that leads to diagonal modulation (i.e. separation between the mean normalized response to matches and the mean normalized response to distractors) involves choosing the linear weights that, once applied, maximize the differences between the variance for the target match normalized responses and the variance for the distractor normalized responses in the linearly transformed responses (e.g. a broad distribution in firing rates for target matches and a narrow distribution for distractors). These variance differences can then be translated into mean differences by a nonmonotonic nonlinearity, such as a squaring operation (see Fig. 6a, center). To find the weights that maximized the variance differences in the normalized responses to target matches and distractors, we designed a method similar in spirit to a PCA (which determines the dimensions with maximal variance). While PCA directly computes the

eigenvectors of the covariance matrix, we first computed the difference between the covariance matrices of the normalized responses to target matches and distractors and then applied the eigenvalue decomposition. The resulting set of eigenvectors thus define the axes along which the variance differences between the normalized responses to target matches and distractors are maximal. Since our task has 16 conditions, the IT population response at a given timepoint has 15 degrees of freedom, i.e. 15 orthogonal axes with a significant amount of modulation of any kind. Because the first model PRH neuron described above captures one degree of freedom, the remaining variance differences are captured by the first 14 eigenvectors described above, and we use these to define the linear weights for the second class of neurons (after reversing the response normalization). To translate any variance differences produced into diagonal modulation, the resulting linear responses were centered (by subtracting the mean μ) and passed through a squaring nonlinearity (Adelson and Bergen 1985) to produce the final response matrix LN_2 :

$$LN_2 = (L - \mu)^2 \tag{19}$$

Finally, the remaining 149 eigenvectors were used to define the linear weights for the third class of model PRH neurons (after reversing the response normalization). While these axes are not required to capture information at the timepoint used to fit the model (see above), they are required to capture all the remaining information that exists within IT at different timepoints (see Fig 6a, right). These linear combinations were exponentiated (Equation 18) to produce final response matrices.

Quantification of code non-stationarities

Code non-stationarities were defined as changes across time in a neuron's response modulations other than rescaling. In our analysis, we measured the degree of similarity of the responses at the reference timepoint of 135 ms and every other timepoint by computing the Pearson's correlation coefficient. To determine the probability that differences arose from noise, we applied a split-half procedure. For each neuron, we began by determining the null-distribution of correlations at the reference timepoint (135 ms) by bootstrapping the correlation across many random split-halves draws across our set of repeated presentations. Next, we applied a similar procedure to compute the test distribution of correlations between the components at 135 ms and those at every other timepoint. The degree of non-stationarity was then measured via a non-parametric comparison between the null and the test distributions (Fig 8). More specifically, we computed the differences between randomly-paired correlation values from the test and the bootstrap distributions, and we measured the p-value as the fraction of instances in which the correlation value across different timepoints was larger than the correlation value for split halves at 135 ms (Efron and Tibshirani 1994).

Pseudosimulation

A pseudosimulation approach was used to determine the contribution of each type of IT non-stationarity on the untangling dynamics of our PRH model. As an overview, we selectively manipulated different features of the noise-corrected IT

modulations to make them stationary, regenerated Poisson trial-by-trial variability, re-applied our PRH model to the modified IT population, and quantified the delays between ideal observer classifier and linear classifier performance (Fig 9). Enforcing stationary responses was accomplished by modifying the structure of neural responses at each timepoint to resemble those at the reference timepoint of 135 ms, but rescaled such that the absolute amount of each signal type did not change. More specifically, we deconstructed the population response at each timepoint into three components: the total amounts of cognitive and visual modulation (i.e. the sum of modulations across the population), the pattern of modulations across neurons (i.e. the extent to which each neuron contributes to the overall modulation), and the code of each neuron's modulation (i.e. the selectivity for each component). In our pseudosimulations, we always maintained the total modulation computed at each timepoint. To measure the impact of modulation non-stationarites, we manipulated each neuron's components to match the code at 135 ms while leaving the modulation pattern intact. To measure the impact of code non-stationarities, we fixed the modulation pattern to match that at 135 ms while allowing the code components to change across time.

As a first step, the 15 bias-corrected modulation components for each neuron were computed for any given timepoint. Only the visual and cognitive (working memory and non-diagonal) components were manipulated. The total amounts of visual ($M_{vis,pop}$) and cognitive modulation ($M_{wm,pop}$ and $M_{cog,pop}$, indicating working memory and the other cognitive components respectively) were always preserved, and the components were normalized by dividing them by the total amount of modulation:

$$m_i' = \frac{m_i}{M_{vis,pop}} \text{ for all visual components;} \quad m_i' = \frac{m_i}{M_{wm,pop}} \text{ for all working memory components;} \quad m_i' = \frac{m_i}{M_{cog,pop}} \text{ for all other cognitive components} \tag{20}$$

The strength of the working memory components and the remaining cognitive components were computed separately to avoid mixing the actual working memory modulations present before the arrival of the visual signals from the spurious noise present at those early timepoints in the remaining modulation components.

To measure the effect of cognitive non-stationarities (Fig. 9c) we preserved the normalized cognitive components measured in our data but replaced the normalized visual components with those measured at 135 ms, followed by rescaling to maintain the total visual modulation at that timepoint. Conversely, to measure the delay due to visual non-stationarities (Fig. 9d) we preserved the normalized visual components but replaced the normalized cognitive components with those measured at 135 ms, followed by rescaling.

To manipulate the modulation non-stationarities in a manner that did not impact the code non-stationarities, we needed to quantify the fractional contribution of each neuron in the population to the overall modulation for each type of signal. We did this by separately summing the squared normalized visual components, the squared normalized working memory components for each neuron and the remaining squared normalized cognitive components, thus resulting in three vectors expressing the relative contribution of each neuron to the total visual modulation and the total cognitive (working memory and remaining cognitive) modulation:

$$v_{vis,n} = \sum_{i \in vis} m_i'^2 \; ; \qquad v_{wm,n} = \sum_{i \in wm} m_i'^2 \; ; \qquad v_{cog,n} = \sum_{i \in cog} m_i'^2 \tag{21}$$

where $v_{vis,n}$ represents the entry for the n-th neuron in the vector expressing the relative contributions to total visual modulation, $v_{wm,n}$ represents the entry for the n-th neuron in the vector expressing the relative contributions to total working memory modulation, and $v_{cog,n}$ represents the entry for the n-th neuron in the vector expressing the relative contributions to total remaining cognitive modulation. Finally, dividing the normalized components by $v_{vis,n}$, $v_{wm,n}$ and $v_{cog,n}$ we obtained a set of "neuron-normalized" components whose values express each neuron's response preferences independent each neuron's relative modulation:

$$m_i'' = \frac{m_i'}{v_{vis,n}} \text{ for all visual components;} \qquad m_i'' = \frac{m_i'}{v_{wm,n}} \text{for all working memory components;}$$

$$m_i'' = \frac{m_i'}{v_{cog,n}} \text{ for all other cognitive components} \tag{22}$$

To measure code non-stationarities (Fig. 9f), we replaced the neuron-normalized components m_i'' with those at 135 ms, followed by rescaling to maintain the total modulation at that timepoint. Conversely, to measure modulation non-stationarities (Fig. 9e) we replaced the vectors v_{vis} , v_{wm} and v_{cog} with those measured at 135 ms (while maintaining the neuron-normalized components m_i''), followed by rescaling.

Results

To explore the neural mechanisms involved in finding visual targets, macaque monkeys performed a well controlled yet simplified version of target search in the form of a delayed-match-to-sample task (Fig 2a) as we recorded neural responses in IT and PRH. On each trial, monkeys sequentially viewed images and indicated when a target image appeared. We held the target fixed in short blocks of trials and we presented the same images as both targets and as distractors in different blocks. Our experimental design included four images in all possible combinations as a visual stimulus ("looking at"), and as a target ("looking for"), resulting in 16 experimental conditions arranged in a four-by-four matrix (Fig. 2b). In these matrices, conditions with a fixed visual stimulus correspond to columns, and conditions with a fixed target (or "working memory") correspond to rows. Additionally, the task required the monkeys to differentiate target match conditions, which fall along the diagonal of this matrix, from distractor conditions, which fall off the diagonal (Fig 2b).

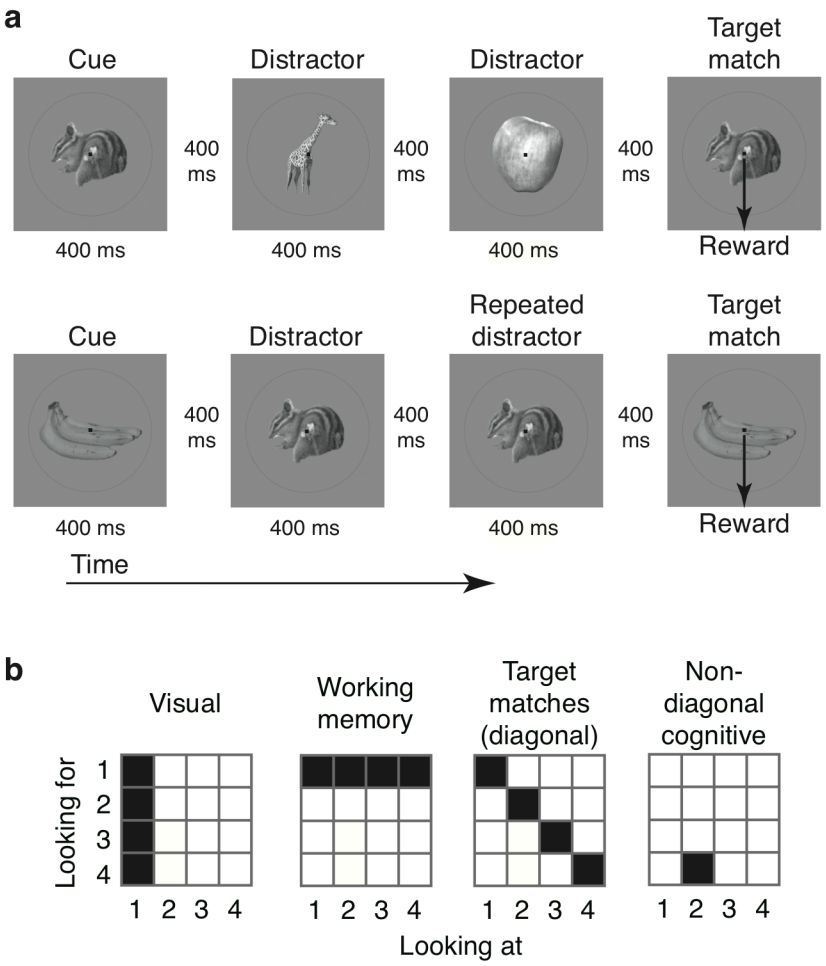

Figure 3-2. *The delayed match-to-sample task.* **a)** Monkeys performed a delayed match-to-sample task that required them to treat the same four images as target matches and as distractors in different blocks of trials. Monkeys initiated a trial by fixating a small dot. After a delay, an image indicating the target was presented, followed by a random number (0-3, uniformly distributed) of distractors, and then the target match. Monkeys were required to maintain fixation throughout the distractors and make a downward saccade when the target appeared to receive a reward. Approximately 25% of trials included the repeated presentation of the same distractor with zero or one intervening distractors of a different identity, similar to Miller and Desimone (1994). **b)** Each of four images were presented in all possible combinations as a visual stimulus ("looking at"), and as a target ("looking for"), resulting in a four-by-four response matrix. In these matrices, conditions corresponding to the same "visual" input correspond to columns, conditions corresponding to the same "working memory" or target input correspond to

rows, and target matches fall along the “diagonal” while distractors fall off the diagonal. The type of matrix structure required to differentiate other types of conditions, (e.g. looking at image 2 and for image 4) are referred to as “non-diagonal cognitive”.

The “untangled” PRH target match representation is initially “tangled”

As described above, computing the solution to the monkeys’ task (i.e. determining whether a currently-viewed image is a target match or a distractor) requires combining visual and working memory information. In a recent paper (Pagan, L.S. et al. 2013), we reported evidence that these signals combine within or before IT in the ventral visual pathway in a largely nonlinearly separable or “tangled” manner (i.e. one in which target match information is present but is not accessible to a linear population read-out; Fig 1, bottom right), followed by computations in PRH that reformat this information into a more linearly separable or “untangled” format (i.e. one more accessible to a linear population read-out; Fig 1, top right). This evidence was based in part on finding similar amounts of “total” target match information in IT and PRH, as measured by the performance of an ideal observer (see Methods, Equations 2-5), while also finding that a larger portion of this information was “linearly separable” (or “untangled”) in PRH as compared to IT, as measured by the performance of a linear classifier applied to the same data (see Methods, Equation 1). To gain deeper insight into the computations implemented by PRH to reformat nonlinearly separable target match signals arriving from IT, we investigated the temporal dynamics with which “total” and “linearly separable” signals evolved. We performed these analyses based on the spike count responses computed in 25 ms windows and we systematically shifted the positions of

the windows relative to the onset of each visual image presented during our experiment. We found that total information arrived in PRH earlier than linearly separable information (Fig 3a-c), consistent with target match information that initially arrived in PRH as partially “tangled”, followed by the arrival of more “untangled” target match information after a short delay.

Quantifying the magnitude of the delay, or equivalently the differences in the latencies with which total versus linearly separable target match information arrived in PRH, required us to set a performance criterion to compute latency (e.g. the time required for performance to reach 0.65). We computed latencies for a range of such criteria (see Methods, Equation 6). We found that the latency differences between total and linearly separable information were fairly constant across the broad range of performance criteria for which we were able to determine them (range: 0.55-0.775; latency difference range: 9.3-12.9 ms, mean latency difference = 11.7 ms; Fig 3b, left) and that these latency differences were significant (e.g. $p=0.011$ for a criterion of 0.65 and $p<0.05$ for all criteria 0.6-0.775). While the latencies of linearly separable signals computed in this manner were longer than those for nonlinearly separable signals, linearly separable signals were also slightly smaller in their overall magnitude (e.g. Fig 3b, left, yellow). To determine the degree to which these magnitude differences produced the latency differences we were observing, we selected the subset of our data in which performance of the linear classifier and ideal observer was matched in a window averaged 135-140 ms, and we recomputed latencies for the same trials at earlier timepoints (see Methods). Average latency differences were similar, albeit slightly smaller for magnitude matched data (mean delay across all criteria was 11.2 ms for the magnitude matched as compared to 11.7 ms for the original data) and magnitude

matched latency differences remained significant across a broad range of criteria (e.g. p=0.018 at a criterion of 0.65; p<0.05 for criterion 0.6-0.775; Fig 3b, right). The delay in the arrival of linearly separable as compared to total information was confirmed in each monkey individually (e.g. for a criterion of 0.65, monkey 1: delay = 15.4 ms, P=0.031; monkey 2: delay = 12.2 ms, P=0.034; not shown; for magnitude matched data, monkey 1: delay = 13.9 ms, P=0.046; monkey 2 delay = 10.9 ms, P=0.048; Fig 3c).

One possible interpretation of these results is that the format of target match information in PRH changes over time from a more nonlinearly separable to a more linearly separable format. However, the analyses described above were performed using two read-out approaches (i.e. an SVM linear classifier and an ideal observer nonlinear classifier) that, while relatively common, differ in their numbers of parameters and how the parameters are optimized. These differences may confound the interpretation of our results, particularly given that above we indirectly infer the amount of nonlinearly separable information at each point in time by comparing "total" and "linear" performance as opposed to measuring it directly. As a control analysis, we developed two new classifiers to measure linear and nonlinear information directly and in a comparable manner, including matched numbers of parameters. Our approach is analogous to a polynomial expansion in that it seeks to deconstruct the classifier decision boundary into a set of terms of increasing order (e.g. $w_1*x + w_2*x^2$ …). To equate the numbers of "linear" and "nonlinear" parameters, the "linear" classifier was characterized by parameters associated with the first-order term (i.e. the means of the firing rate distributions) and the "nonlinear" classifier was characterized by parameters associated with the second-order term (i.e. the variances of the firing rate distributions; see Methods). A comparison of the temporal evolution of performance for this linear and this

nonlinear classifier (Fig 3d, left) revealed that at early times (e.g. 80 ms), linear and nonlinear performance were approximately matched but at later times (e.g. 110 ms), nonlinear performance began to plateau as linear performance continued to rise. Consequently, chance-corrected linear classifier performance grew to nearly 3-fold nonlinear performance by 140 ms (Fig 3d, right). These results are consistent with a target match representation that arrives in PRH as partially tangled (i.e. an approximately balanced combination of nonlinear and linear target match information) and then becomes more untangled (i.e. more linearly separable). These results are inconsistent with the alternative proposal that target match information increases in its overall amount but does not change its format with time.

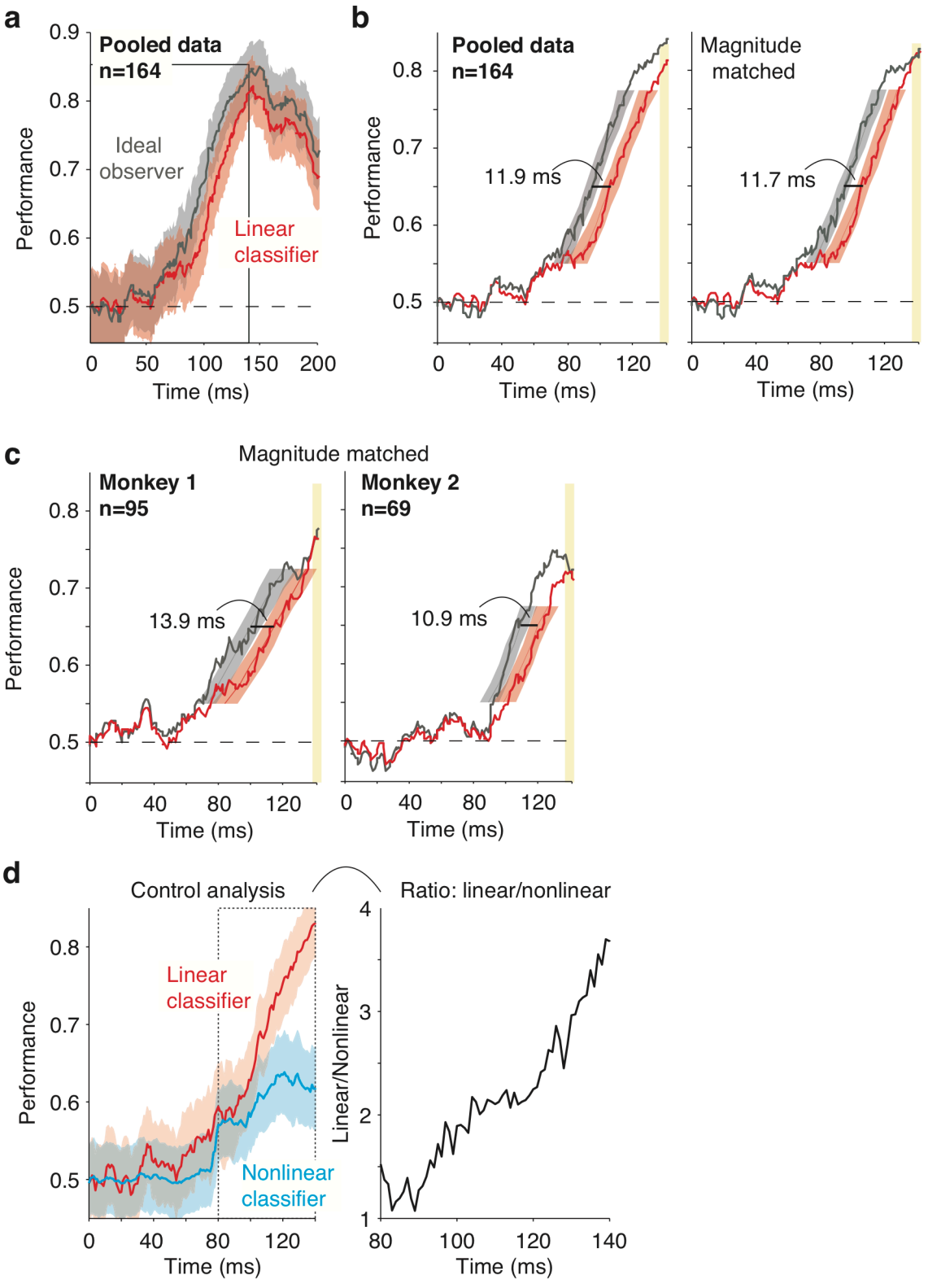
a
Pooled data
n=164
Ideal
observer
Linear
classifier
Performance
Time (ms)
b
Pooled data
n=164
11.9 ms
Magnitude
matched
11.7 ms
Performance
Time (ms)
c
Magnitude matched
Monkey 1
n=95
13.9 ms
Monkey 2
n=69
10.9 ms
Performance
Time (ms)
d
Control analysis
Linear
classifier
Nonlinear
classifier
Performance
Time (ms)
Ratio: linear/nonlinear
Linear/Nonlinear
Time (ms)

Figure 3-3. *Target match signals are gradually untangled in PRH.* Comparison of the temporal evolution of the performance of an ideal observer classifier, to assess the amount of total target match information, and a linear classifier, to assess the amount of target match information that was accessible to a linear read-out. Cross-validated performance was computed with spike counts within 25-ms bins sampled at 1-ms intervals. In all panels, the horizontal dotted line indicates chance performance and "n" indicates the number of neurons included in each population. **a)** Performance for the data pooled across both monkeys; shaded region indicates standard error of performance (the y-axis); **b)** *left,* the same data in panel a shown from 0 – 140 ms to more closely examine the delay but with standard error computed for time (the x-axis), *right,* the same analysis performed on a subset of trials in which performances were matched in magnitude from 135-140 ms (see Methods); **c)** the same analysis presented in panel b, right, but applied to the data from each monkey individually; **d)** *left,* as a control analysis, direct measures of linear and nonlinear performance using two classifiers matched for numbers of parameters (see Methods); *right,* the ratio of chance corrected linear and nonlinear performance computed from the plots on the right where the ratio was determined for each time bin as (linear – 0.5)/(nonlinear – 0.5).

What types of single-neuron responses account for population untangling dynamics?

The population-based framework described above is useful for understanding the combined PRH population representation. As a complementary analysis, we were interested in relating these descriptions of population dynamics with more intuitive descriptions of the signals reflected in the responses of individual neurons. As an overview of how we determined this relationship, we applied a technique to parse each neuron's responses into intuitive components (i.e. the magnitudes of visual and different

types of cognitive modulation) and we derived the relationships between these single-neuron modulations and population performance for the ideal observer and linear classifiers. Our decompositions assume that population performance is not impacted by correlated trial-by-trial variability between neurons ("noise correlations"), which we have previously determined to be true in our data (Pagan, L.S. et al. 2013).

To deconstruct the firing rate modulations of each neuron into intuitive components, we applied a noise-corrected, ANOVA-like analysis (see Methods, Equations 7-9) to parse each neuron's responses into firing rate modulations that could be attributed to: changing the visual image ("visual"; Fig 2b); changing the identity of the target ("working memory"; Fig 2b); changing whether a condition was a target match or a distractor ("diagonal"; Fig 2b); and changes between other "non-diagonal cognitive" conditions (e.g. looking at image 2 and for image 4 versus for image 3; Fig 2b). Figure 4a shows the decomposition for three example neurons and Figure 4b shows the total magnitudes of these signals across the PRH population as a function of time. We found that the visual signal was the strongest type of signal in PRH, followed by the diagonal, working memory and non-diagonal signals, respectively. Consistent with weak "persistent activity", working memory signals were present before the other signals, which followed a stimulus-evoked timecourse.

Next we determined the relationship between the magnitudes of these signals and the predicted population performance for both read-outs as a two-stage process in which signals were first combined into "classifier components" which were then converted to performance values via a mapping function (Fig 4b; see Methods, Equations 11,12). We imposed that the mapping function be matched for the two types

of classifiers (i.e. the complementary error function; see Methods) but allowed the two classifiers to rely on different signals. We found that the evolution of linear classifier performance was well-described by the amount of diagonal signal alone (i.e. the "linear component"; Fig 4c, d red). In contrast, we found that accounting for the evolution of ideal observer performance required summing the linear component with a "nonlinear component" term that nonlinearly combined the visual signal and the other two types of cognitive signals (working memory plus non-diagonal cognitive; Fig 4d, cyan; see Methods, Equations 13-15). This result can be understood in the context that performance of the ideal observer depends upon the degree to which the responses to the same images presented as target matches and distractors are non-overlapping (Fig 1) and any type of cognitive modulation (diagonal, working memory, or non-diagonal cognitive) will be at least partially effective at producing this separation. Notably, the temporal dynamics of the linear and nonlinear classifier components inferred from these underlying signals provided a reasonable match to direct measures of the same quantities, including the saturation of the nonlinear component at ~110 ms as the linear component continued to rise (compare Fig 4d, left with Fig 3d), leading to a increasing ratio between linear and nonlinear dynamics as a function of time (Fig 4d, right). This correspondence allowed us to pinpoint the source of the population dynamics in PRH. We found that the saturation of the nonlinear classifier component could largely be attributed to a visual signal that peaked at ~110 ms and then began to fall (Fig 4b grey, Fig 4d cyan, dashed line). In contrast, the linear classifier component continued to rise beyond 110 ms due to a continually rising underlying diagonal signal (Fig 4b, Fig 4d, dashed line). Consequently, the representation of target match signals initially arrived in PRH in a more "tangled" format because visual and working memory signals initially

arrive in PRH to some degree as separate signals (coinciding with an initial wave of diagonal signal) followed by the emergence of a more “untangled” target match representation ~10-15 ms later when diagonal signals become stronger and visual information decreases.

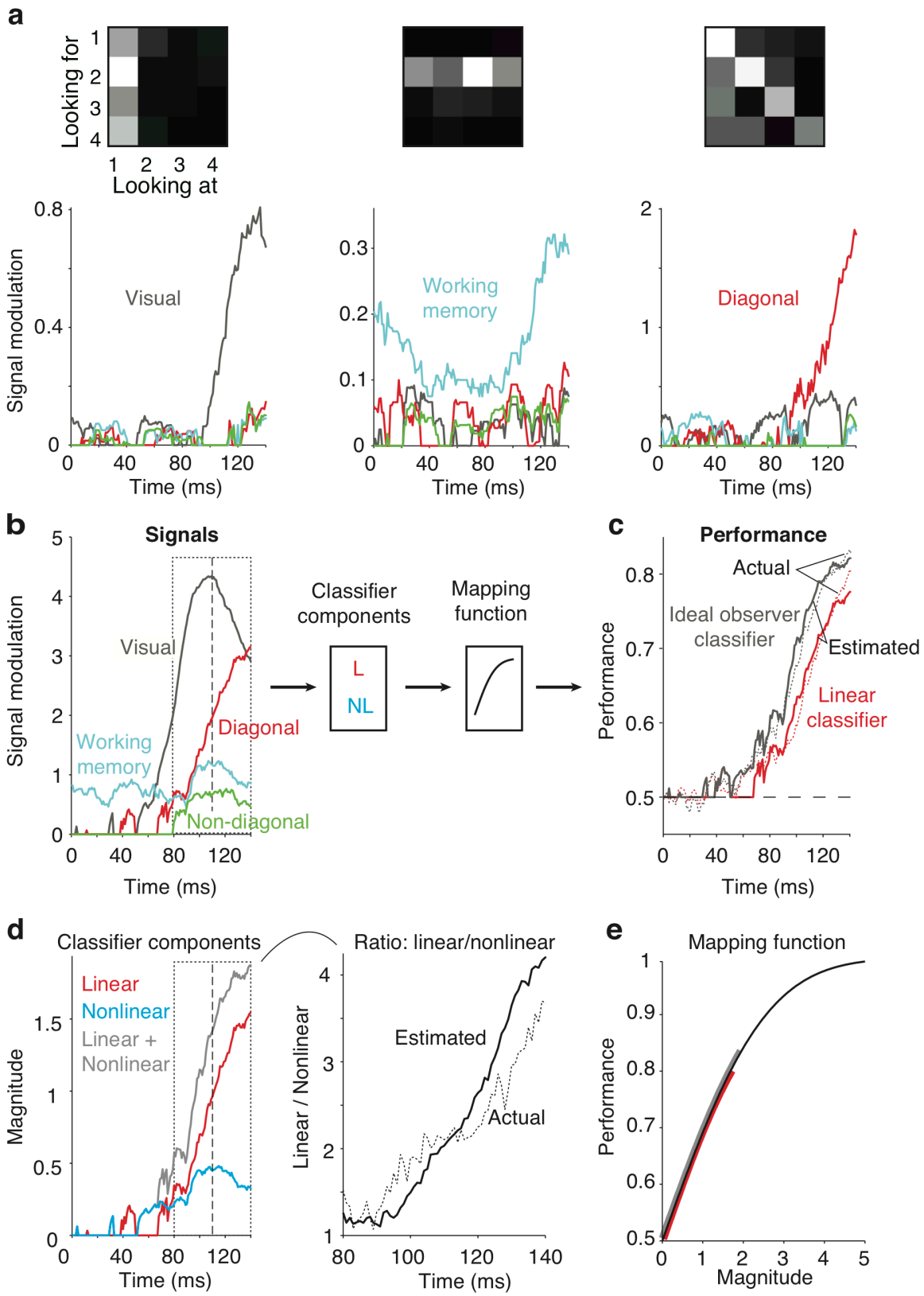
a
Looking for
1
2
3
4
1 2 3 4
Looking at
Signal modulation
Visual
Working memory
Diagonal
Time (ms)
b
Signals
Visual
Working memory
Diagonal
Non-diagonal
Classifier components
L
NL
Mapping function
c
Performance
Actual
Ideal observer classifier
Estimated
Linear classifier
Performance
d
Classifier components
Linear
Nonlinear
Linear + Nonlinear
Magnitude
Ratio: linear/nonlinear
Linear / Nonlinear
Estimated
Actual
e
Mapping function
Performance
Magnitude

Figure 3-4. *Single-neuron decomposition of population untangling dynamics in PRH.* To determine the relationship between single-neuron response properties and population performance measures, we applied a method to parse each neuron's responses into intuitive signal modulation components, including firing rate modulations that could be attributed to: "visual" - changes in the visual image; "working memory" - changes in the identity of the target; "diagonal" - whether a condition was a target match or distractor; "non-diagonal" - other cognitive modulations (see Methods, Equations 7,8). We note that the method estimates and corrects for noise to ensure that trial-by-trial variability is not confused with signal. **a)** Plotted is the strength of each type of modulation (Equation 8) as a function of time relative to stimulus onset, for three example neurons whose responses are dominated by one type of modulation. Also shown are the firing rate response matrices for each neuron, with spike counts averaged within the same window (0-140 ms after stimulus onset), each rescaled from the minimum (black) to maximum (white) firing rate. **b)** *Left,* The same plots depicted in panel a, but summed over all neurons in the PRH population. *Right,* The relationship between these signal modulation magnitudes and performances for the ideal observer and linear classifier can be described as a "classifier component" computed from the underlying signals followed by a mapping function that transforms the component values into performances. The classifier component for the linear classifier was computed from the diagonal signal alone. The classifier component for the ideal observer was computed by summing the linear classifier component signal with a second "nonlinear classifier component" that nonlinearly combined the visual and other cognitive signals (working memory and non-diagonal cognitive; Methods, Equations 13-15). The same mapping function was used for both classifier predictions. **c)** Timecourses for the actual (dotted, replotted from Fig 3b, left) and estimated (solid) classifier performance values. **d)** Timecourses of the linear, nonlinear and summed classifier component signals. **e)** Timecourse for the actual (dotted, replotted from Fig 3d) and estimated (solid, based on the data on the left) ratios of chance corrected linear and nonlinear classifier components with the same conventions as Fig 3d, right. **f)** The mapping function used to convert classifier component magnitudes into performance predictions. The red and gray lines indicate the range of values used to estimate the linear and ideal observer, respectively. In

panels b and d, dotted box from 80 – 140 ms and the dashed line a 110 ms are provided as visual benchmarks.

Dynamic representation in PRH can be accounted for by instantaneous PRH computation

The dynamics underlying untangling in PRH could provide an important constraint on descriptions of how this computation is implemented, and thus we were interested in determining the classes of models that could account for the delay between nonlinearly separable and linearly separable target match information in PRH. Following from our previous results (Pagan, L.S. et al. 2013), we can begin by ruling out simple descriptions in which these delays are entirely inherited from the primary input to PRH, IT, because IT contains less linearly separable information. As illustrated in Fig 5a-c, the linearly separable target match information (and corresponding diagonal signals) that do exist in IT are delayed relative to total information (and other types of signals) but they are smaller in magnitude than those in PRH. Thus while the delays between the arrival of "tangled" versus "untangled" information in PRH are likely inherited in part from IT, they cannot fully account for the result.

In our previous report, we presented evidence that a simple, feed-forward model could account for the transformation of other types of IT signals into diagonal signals in PRH (Pagan, L.S. et al. 2013). In that model, computation in PRH was instantaneous (and spike count windows were broad). Thus upon finding that target match signals evolved dynamically in PRH, we naturally assumed that accounting for these dynamics

would require us to extend our model to incorporate dynamic PRH computation (e.g. as a result of implementing these computations in complex, recurrent circuits). We were very surprised to discover that instead of attributing these delays to PRH, they could be accounted for by a variant of a feed-forward model in which PRH computations were fixed and acted instantaneously - but crucially - upon input from IT that changed its content over time (as described below).

To evaluate this class of model, we considered whether a model fit to our recorded IT responses could produce a model population that reproduced the dynamics we observed in PRH. To constrain the fits, we assumed that computations in PRH sought to transform the maximal amount of nonlinearly separable (i.e. “tangled”) information arriving from IT into a linearly separable (i.e. “untangled”) format. Fitting an instantaneous, feed-forward model of PRH computation that maximally extracted diagonal signal from our recorded IT responses required us to develop novel model fitting procedures. The novel model fitting procedures we describe here incorporate non-trivial extensions to ones we have previously reported (Pagan, L.S. et al. 2013).

As an overview, the responses of each model PRH neuron were computed via a linear-nonlinear (LN) model as a weighted combination of all IT neurons, followed by an instantaneous nonlinearity. The input to the model consisted of the responses of 164 IT neurons to the 16 experimental conditions and the output of the model consisted of 164 model PRH responses to those same conditions. Model responses were determined for individual trials (i.e. trial-by-trial variability in our model PRH was inherited from the recorded IT responses) and the model was fully cross-validated, meaning that we used 50% of our data to train the model (10 repeats for each condition) and we assessed

model performance using the other half of our data (the other 10 repeats). As described in more detail below, we fit the model to the IT responses at a single timepoint (e.g. 135 ms) and these parameters were held fixed for all the other timepoints. We emphasize that the model fits are confined to the data recorded from IT and our goal is to evaluate the degree to which these model response properties are similar to our recorded PRH data.

Linearly separable target match information amounts to a difference in the average responses across the set of target matches as compared to the set of distractors (Fig 6a, red versus gray). To understand how the model converted nonlinearly separable target match information into a linearly separable format (i.e. increased mean differences), it is useful to consider the model PRH neurons as three classes (Fig 6a). First, a single model PRH neuron served to combine and inherit all of the linearly separable information that already existed in IT (Fig 6a, left). We determined the linear weights for this neuron (i.e. the weights to apply to each IT input neuron before summation) as the optimal linear target match / distractor discriminant (see Methods). The responses for this neuron were then computed by applying a exponential nonlinearity (e.g. a "soft threshold"; Pillow, Shlens et al. 2008) to the linearly weighted IT responses. The second class of model PRH neurons “computed” linearly separable information from the inputs arriving from IT. We determined the weights for these model neurons using an insight from our previous work (Pagan et al., 2013), in which we established that the crucial property that needs to be maximized through linear weighting is the difference between the variance across the responses to target matches and distractors in the linearly combined responses (i.e. a high-variance for the responses to one set and a low variance for the other; Fig 6a, center). Under appropriate conditions,

these variance differences are transformed into mean differences via a non-monotonic (i.e. a squaring) nonlinearity. We determined the linear weights for these neurons using a method similar to PCA (i.e. an eigenvector decomposition of the difference of the covariance matrices for matches and distractors; see Methods). The number of these neurons was set as the number required to capture all the available information at the training timepoint, given the number of degrees of freedom in our experiment (see Methods). The final class of model PRH neurons served to capture any remaining information at different timepoints (see Methods).

As described above, the goal of our model was to transform the maximal amount of nonlinearly separable (i.e. "tangled") information arriving from IT into a linearly separable (i.e. "untangled") format within the class of models we were working with (i.e. the n-wise LN model). Because the representation arriving from IT changes its content with time (as elaborated below), this required us to select a specific time window for the optimization. To do so, we began by making the reasonable assumption that connectivity between IT and PRH was established for this task during the experience of looking for targets, and we thus looked to the learning literature to guide our selections. We selected the width of our spike count window, 25 ms, to fall within the range of integration times over which synaptic plasticity is thought to occur (Froemke and Dan 2002). We then explored different positions for this window relative to stimulus onset. Training windows placed shortly after total information arrives in PRH, at 75 ms, produced linearly separable target match information without a delay relative to the arrival of total information, but only increased linear classifier performance by a small amount (i.e. this model accounted for 35% of the performance increases observed in PRH over IT; Fig 6b). This was because the parameters that were optimal for these early

time windows failed to generalize to later timepoints where the IT representation differs (described in more detail below). Training windows placed later, at 135 ms, produced larger overall increases in linearly separable information (i.e. this model accounted for 86% of the increases observed in PRH over IT; Fig 6c). However, this information was delayed relative to the arrival of total information. This is because the parameters appropriate for extracting linearly separable information at these later times failed to generalize to earlier times. These results suggest that processing speed and information content trade-off one another in PRH due to the dynamic nature of input arriving from IT.

Strikingly, the model PRH produced by training at the later timepoint of 135 ms had many response properties similar to the actual PRH. Most notably, the magnitude of the delay between the arrival of total and linearly separable information was similar between the model and the actual PRH (for a criterion of 0.65, actual PRH = 11.9 ms, model = 13.8 ms; Fig 6c). The model also approximately reproduced many other notable and subtle properties of the actual PRH responses that were also not directly fit during the optimization. These include the approximate amounts and timecourses of the increases in diagonal modulation from IT to PRH (Fig 6d, red), the decreases in visual modulation from IT to PRH (Fig 6d, gray), and the existence of the working memory modulation before the stimulus evoked response ("persistent activity"; Fig 6d cyan). We note that slightly lower overall ideal observer and linear classifier performance in the model as compared to the actual PRH (Fig 6c, red and gray solid thick versus thin) is partially imposed by slightly lower ideal observer performance in the input to the model, IT (Fig 6c, gray dotted) coupled with the constraint that the model cannot artificially create information.

We emphasize that in our model, delays in the arrival of linearly separable relative to total target match information result in large part from computations that are implemented in PRH instantaneously (i.e. without a delay). How could this be possible? Our model works as follows. Shortly after the onset of a test stimulus (~25 ms; Fig 6d gray dotted), visual signals have not yet arrived in IT and working memory signals exist in isolation; these working memory signals are passed on to PRH. Because computing the target match signal requires both visual and cognitive signals, total target match information is absent in both IT and PRH at this time. Some time later (~75 ms; Fig 6d gray dashed), stimulus-evoked visual signals arrive in IT, which in turn passes both visual and cognitive information to PRH, and total target match information is present in both areas. However, little diagonal signal is created in PRH because the specific contents of the visual and cognitive inputs arriving from IT are misaligned with the biophysical parameters of the PRH neurons (i.e. the synaptic weights), which have been optimized to produce diagonal signals at a later time. Consequently, little untangled target match information exists in PRH. As time passes (~135 ms; Fig 6d gray solid), the specific content of the IT representation becomes aligned to the fixed PRH biophysical parameters (as elaborated below), and diagonal signals are created, thus producing a more linearly separable, untangled target match representation in PRH.

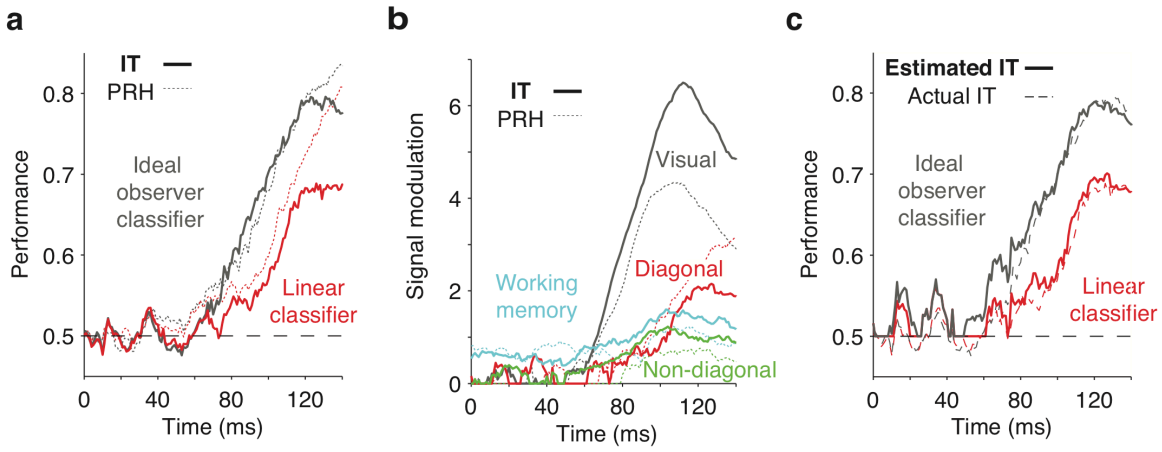

Figure 3-5. *Quantifying population performance and its single-neuron correlates in IT.* **a)** The timecourse of ideal observer and linear classifier population performance in IT (solid), and for comparison PRH (dotted), plotted with the same conventions as Fig 3a. **b)** The timecourse of signal modulation components in IT (solid) and for comparison PRH (dotted), plotted with the same conventions as Fig 4b. **c)** A comparison of actual (dashed) and estimated (solid) ideal observer and linear classifier population performance for IT, plotted with the same conventions as Fig 4c. For both IT and PRH, populations included 164 neurons.

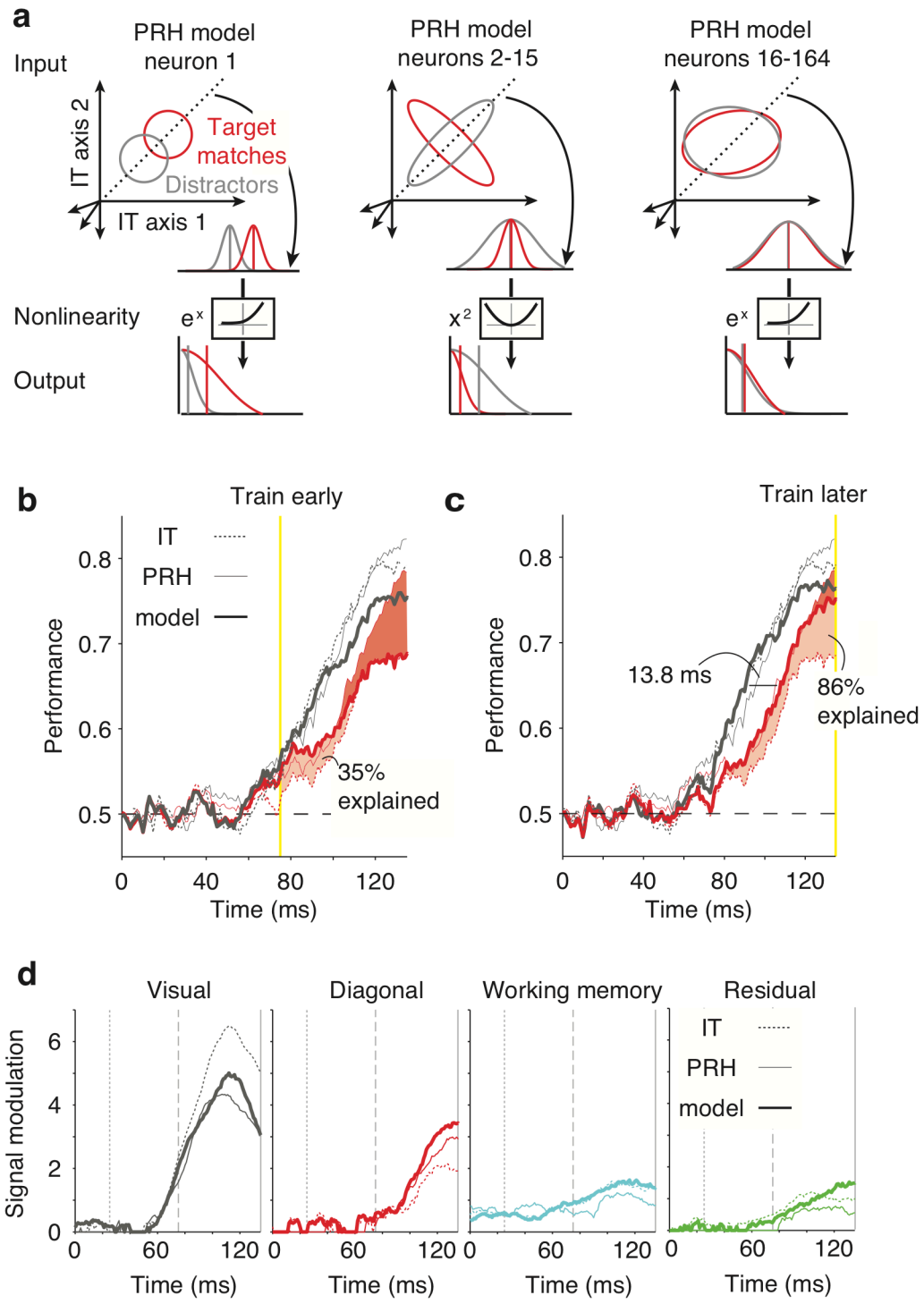

Figure 3-6. *A fixed, instantaneous model of PRH can reproduce the dynamics observed in PRH.* An instantaneous linear-nonlinear model of PRH was fit to maximally untangle the responses of IT neurons. **a)** Three classes of neurons were created to produce the

model PRH population. Shown are idealized depictions of one neuron from each class. For all three classes, the top of each plot ("Input") depicts the hypothetical responses to the set of all target matches (red) and distractors (gray) in 2 dimensions of the 164 dimensional input population space; dotted lines represent the axis along which IT inputs are projected (i.e. the linear weights for one model neuron). Curved arrows point to the distributions of target matches and distractors following weighted linear combination. Below, the same distributions are shown as "Output", following application of a nonlinearity (labeled). The first model neuron (left) inherited all of the linearly separable information available in IT; the linear weights for this neuron were determined as the optimal linear discriminant (i.e. the axis, represented by the dotted line, that maximizes the mean separation for the set of matches from the set of distractors) and the nonlinearity for this neuron consisted of exponentiation. The second class of neurons (2-15; center) "computed" linearly separable information; the weights for these neurons were determined as those that maximized the difference between the variances for target matches and distractors (see Methods) and the nonlinearities consisted of squaring. The final class of neurons (16-164; right) were not required to capture information at the timepoint used to train the model but were required to capture information at other times (see Methods). The linear weights for these neurons were determined as the set of axes that were orthogonal to the previously defined model neurons and those that were necessary to span the remaining IT space and the nonlinearities for these neurons were exponential functions. **b-c)** Timecourse of ideal observer and linear classifier performance when model parameters were optimized for IT responses measured within a 25-ms time window centered at **b)** 75 ms and **c)** 135 ms (yellow lines). Performance is shown for: the model (solid thick), IT (dotted), and PRH (solid thin) for the ideal observer (gray) and linear classifier (red). The increase in linear classifier performance from IT to PRH is indicated as the shaded region, where light red indicates the increases reproduced by each model and darker red indicates the increases that remain unaccounted for; the overall magnitudes of increases that are accounted for are labeled. **d)** Signal types, shown with the same conventions as Fig 4b, for: the model shown in panel c (solid thick), the actual IT data (dotted) and the actual PRH data (solid thin). The gray lines are provided as visual aids to compare the responses at 25 ms (dotted), 75 ms (dashed), and at 135 ms (solid) after stimulus onset.

The IT representation exhibits many different types of non-stationarities

The results presented above suggest that the delays between the arrival of "total" and "linearly separable" target match information in PRH must somehow arise from computations performed on an input representation from IT that changes its content over time (i.e. is "non-stationary") and thus we wished to better understand the specific types of non-stationarities that existed in IT. One useful albeit broad definition of "non-stationarity" is any change in the neural population response other than an overall rescaling. In fact, if the IT modulations were simply rescaled at different points in time (i.e. relative to 135 ms), our model of PRH would not exhibit delays between total and linearly separable information (as shown in Fig 9b). More narrowly, non-stationarities can arise from two conceptually distinct factors. First, non-stationarities can arise from changes in the distribution of information across the neural population over time ("modulation non-stationarities"). For example, information can be carried by different subsets of neurons at different times, due to variability in the response latencies across a population. Consequently, the synaptic strengths appropriate for creating an untangled target match signal at one timepoint can fail to generalize to other timepoints in which different IT neurons carry information (Fig 7a). Second, non-stationarities can arise from changes in the selectivity of individual neurons for the specific components that combine to form the overall modulation envelope ("code non-stationarities"). Similar to the potential impact of modulation non-stationarities, the potential impact of code non-stationarities is an inability to generalize the biophysical parameters that are appropriate

for creating an untangled target match signal at one timepoint to different points in time (Fig 7b).

Measuring code non-stationarities for different types of signals (e.g. visual versus cognitive) required us to develop a way to measure the rank-order selectivity for different "visual" versus "cognitive" components of the signal at different points in time. We note that this cannot be achieved by simply measuring the rank-order selectivity preferences for the 16 different experimental conditions because each condition is a combination of both visual and cognitive information (i.e. a combination of the current visual stimulus and the current target). To parse these signals, we developed a method to linearly transform each neuron's 16 entry response matrix into 16 different "component" responses where 3 of the components describe the visual response, 3 components describe the working memory response, 1 component describes the diagonal response, 8 components describe the non-diagonal cognitive responses, and a final component corresponds to the neuron's grand mean firing rate (Fig 7b, bottom; Methods Equations 7,8). Together, these components form an orthonormal basis and thus this procedure is similar to a PCA but instead of finding the stimulus dimensions that account for the most variance, we assign the dimensions *a priori* to capture intuitive, task-relevant components of a neuron's response, and determine the amount of firing rate modulation along each dimension (Methods, Equations 8,9). Notably, the different components combine to form the signal modulation envelopes depicted in Fig 4a-b (e.g. the 3 visual components combine to determine the visual signal modulation envelope; Methods, Equation 8).

Figure 8 includes a visualization of IT component and code non-stationarities (similar to Brincat and Connor, 2006) analyzed separately for the cognitive (Fig. 8a) and visual (Fig. 8b) signals. In these plots, rows correspond to the responses of individual neurons, plotted as a function of time relative to stimulus onset. The modulation envelope for each neuron is depicted by brightness (black to bright), and neurons are ranked by the times at which the peaks of their envelopes fell. Modulation non-stationarities are thus indicated by changes in the brightness patterns between two columns of the plot. As illustrated by the considerable change in the subpopulations of neurons that were active at (e.g.) 75 versus 135 ms (orange lines), modulation non-stationarities were present in both the visual and cognitive signals in IT. This can also be seen by examining the modulation envelopes for four example neurons with a variety of latencies and peak response times (Fig 8a-b, left).

In contrast, the degree of code non-stationarity for each neuron (relative to 135 ms) is indicated in these plots by color, with stationary responses indicated in yellow and non-stationarities indicated in blue. To measure code non-stationarities, we compared each neuron's selectivities for the different components at 135 ms with its selectivities at every other timepoint, and we determined the probability (the p-value) that changes in selectivity were due to trial-by-trial variability (see Methods). As illustrated by the presence of blue in these plots, (Fig 8a-b, center), code non-stationarities were present in both the cognitive and the visual signals. Example neurons with visual and cognitive codes that were both stationary and non-stationary are shown (Fig 8a-b, right).

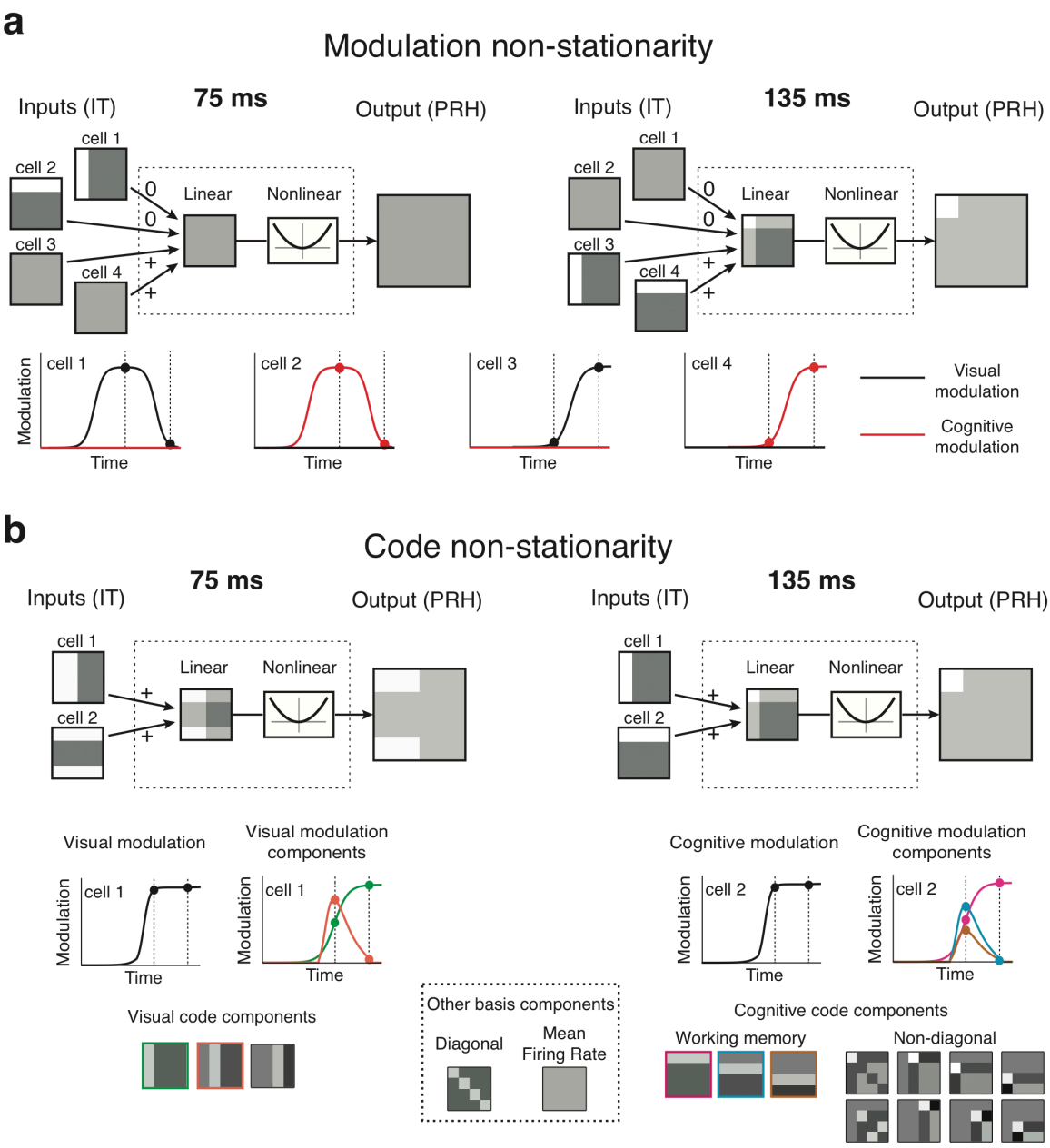

Figure 3-7. *The hypothetical impact of IT modulation and code non-stationarities on PRH computation.* **a)** A hypothetical illustration of how modulation non-stationarites in IT could impact computation in PRH. Two pairs of hypothetical IT cells (cells 1 and 2 versus 3 and 4) are shown, each which contain one visual and one working memory neuron. Shown at the bottom are plots of the visual (black) and cognitive (red) modulation magnitudes as a function of time. Note that the two pairs are maximally activated at different times (e.g. 75 versus 135 ms). Thus a model fit to extract diagonal signal at

135 ms (resulting in positive weights on neurons 3 and 4) will not generalize to produce diagonal signals at 75 ms. **b)** A hypothetical illustration of how code non-stationarities in IT could impact computation in PRH. Shown are one visual and one working memory neuron that combine to form a diagonal signal at 135 ms. While the modulation magnitudes (i.e. the envelope of the combined visual and cognitive signals) of these hypothetical neurons are matched at 75 and 135 ms, the code (i.e. the response selectivity for the different visual and cognitive components) differs between these two timepoints, and thus a model fit at 135 ms will not generalize to produce diagonal signal at 75 ms. Visual and cognitive code components were computed by decomposing the response matrix at each timepoint via the linear basis (shown at the bottom; Methods, Equation 7).

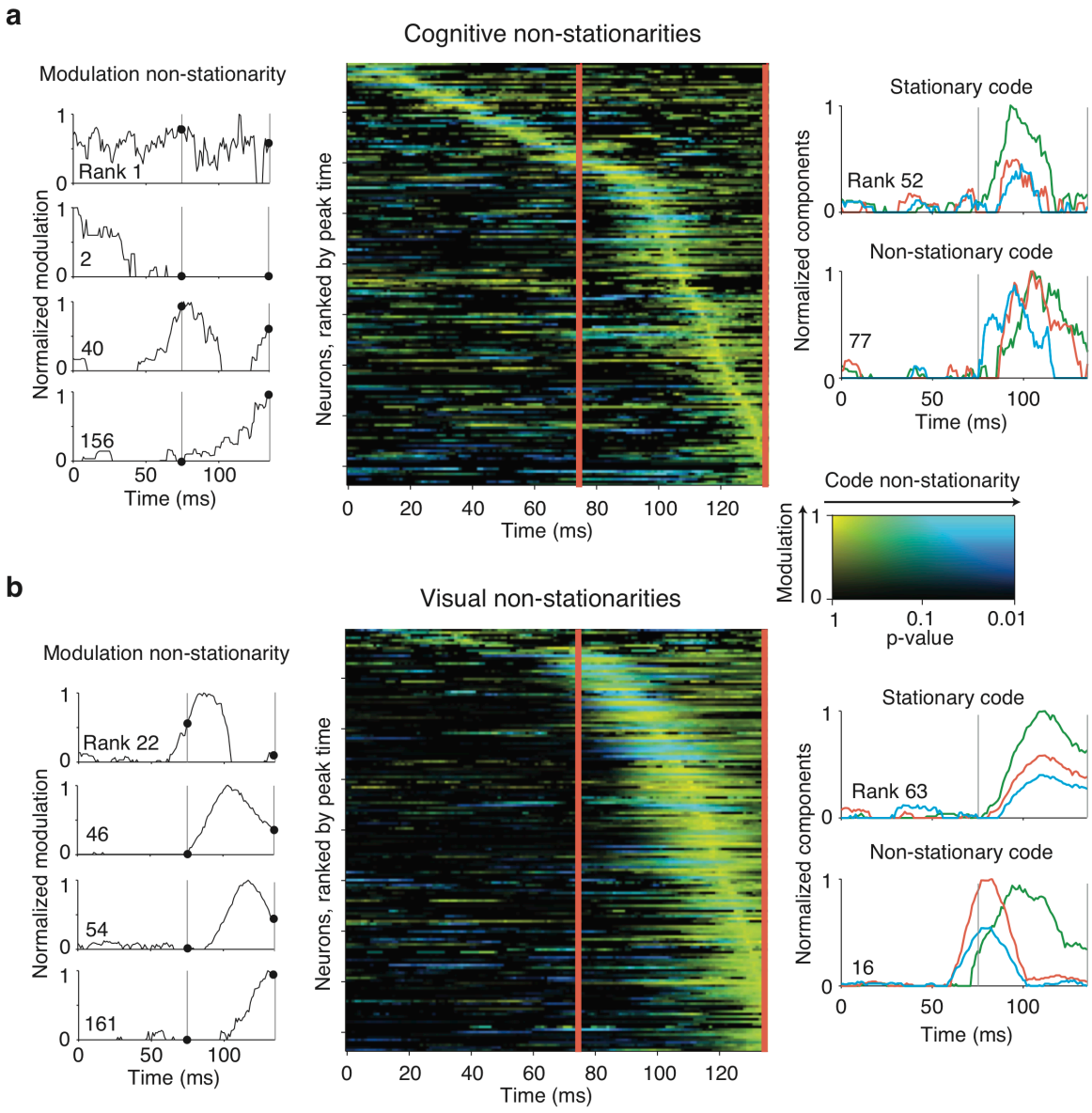

Figure 3-8. *Visual and cognitive non-stationarities in IT.* **a)** Quantification of the cognitive non-stationarities in our IT data. *Center*, Each row represents one neuron's cognitive non-stationarity as a function of time relative to stimulus onset; neurons are ranked by the peak time of their cognitive modulation. Brightness indicates the magnitude of cognitive modulation (i.e. the envelope of the combined cognitive signal for the 3 working memory and 8 non-diagonal cognitive components), relative to each neuron's peak, while hue indicates the degree of cognitive code non-stationarity (i.e. changes in the selectivity for different cognitive components) relative to the 135 ms timepoint, with

stationary responses in yellow and non-stationarities in blue. Code components were computed at each timepoint as described in Fig 7b and methods. The degree of code non-stationarity was measured by a noise-adjusted, cross-validated analysis which quantified the probability (the p-value) that changes in the code between two timepoints arose from noise (see Methods). *Left*, plots of cognitive modulation as a function of time, normalized to range from 0 to 1, for four example neurons. *Right*, plots of the largest 3 (of 11) cognitive code components as a function of time, for two example IT neurons. **b)** Visual non-stationarities, plotted using the same conventions as in panel a. In all plots, vertical lines are provided as visual aids to compare responses at 75 ms and at 135 ms.

Code non-stationarities in IT are the largest contributors to PRH model dynamics

The analysis presented above suggests that many different types of non-stationarities exist in IT (i.e. modulation and code non-stationarities for both visual and cognitive signals); to what degree did the dynamics of our PRH model depend on each type? To evaluate this question, we performed a series of pseudosimulations in which we manipulated our recorded IT responses such that one or more types of signals were artificially made stationary, and we quantified the delays that remained between total and linearly separable information in our model of PRH. For example, to quantify delays due to modulation non-stationarities, we manipulated the data such that the selectivity to code components for all neurons was perfectly stationary relative to the 135 ms timepoint used to train the model, while preserving any modulation non-stationarities that existed in the data (see Methods, Equations 20-22). Similarly, to quantify the delays due to code non-stationarities, we enforced the modulation signals to be perfectly stationary by adjusting the relative contribution of each neuron (i.e. the magnitude of the “envelope”

for each type of modulation) at every timeslice to match the 135 ms reference timeslice, while preserving any code non-stationarities that existed in the data (see Methods, Equations 20-22). Notably, the impact of both types of pseudosimulations was confined to the format of the signal components (by changing their distribution across neurons or the code selectivity within individual neurons), and never modified the total amount of any type of signal modulation across the population at any timeslice. The results of these simulations revealed that many different types of non-stationarities contribute to the delays between the arrival of total and linearly separable information in our model of PRH, with the cognitive and code non-stationarities being the largest contributors (Fig 9). The fact that visual non-stationarities exist in IT (Fig 8b) but do not provide a sizable contribution to the delays we observe in our model (Fig 9d) can be explained by the fact that ideal observer performance relies on a combination of visual and cognitive signals (Equations 13-15) and because cognitive signals are smaller, their non-stationarities play a larger role (e.g. they serve as a “bottleneck” for model computation).

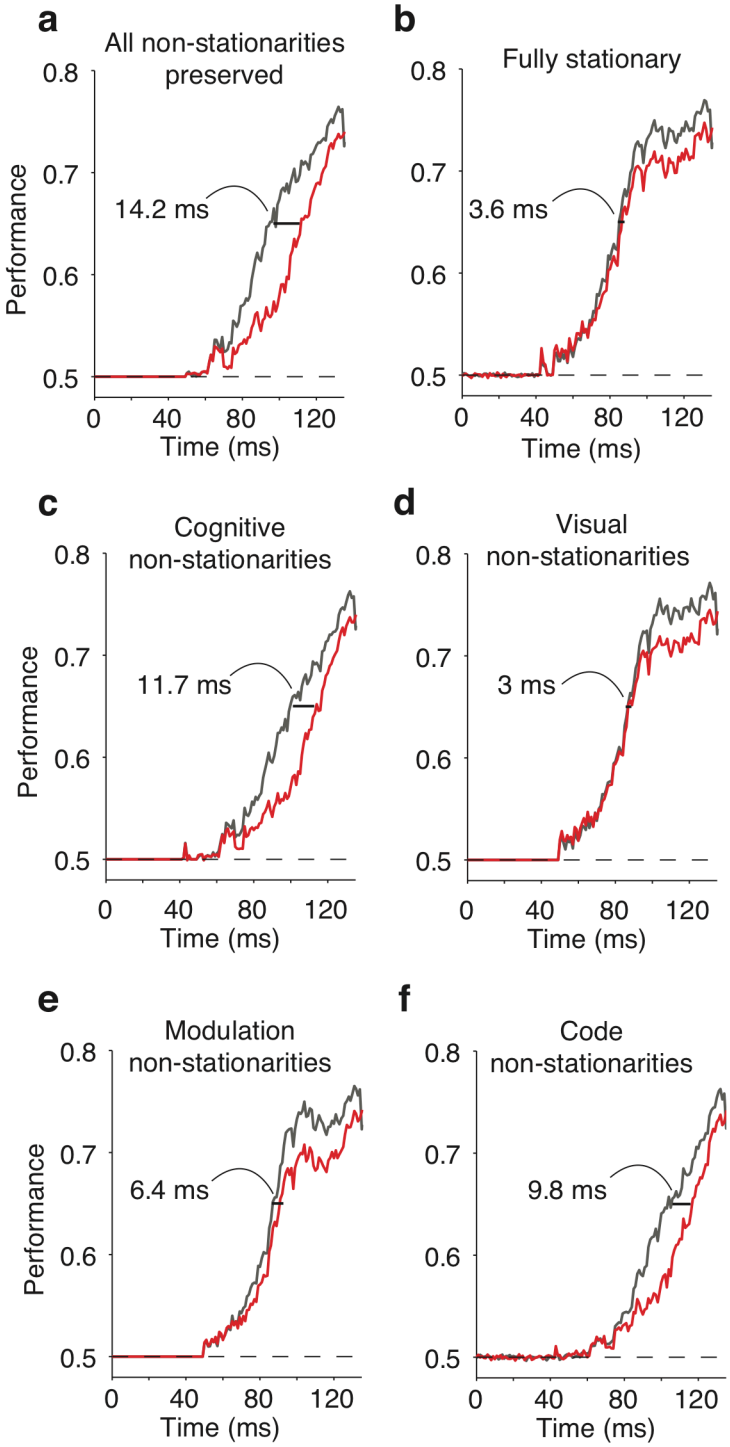

Figure 3-9. *Impact of IT non-stationarities on the untangling dynamics of a model of PRH.* To determine the effect that different types of IT non-stationarities might have on the dynamics of the target match representation in PRH, we performed a series of

pseudosimulations in which we selectively imposed that one or more types of IT signals were perfectly stationary while leaving the others untouched, and we measured the delay that remained between ideal observer and linear classifier performances computed from the manipulated model PRH responses. **a)** The PRH model with no signal manipulation, but with Poisson trial-by-trial variability regenerated for IT (see Methods; compare with the actual data in Fig 3). **b)** Manipulating all IT signals to become stationary nearly eliminates the delay in PRH. **c-f)** Contribution of the following types of IT non-stationarities to the delay observed in the PRH model, measured by making all other types of signals stationary: **c)** cognitive (both code and modulation), **d)** visual (both code and modulation), **e)** modulation (both visual and cognitive) **f)** code (both visual and cognitive).

Discussion

One of the biggest challenges in studying a high-level brain area like PRH is parsing the response properties that have been "inherited" from its inputs from those response properties that are "computed" at that stage, and our findings demonstrate that these determinations need to be made carefully. Here we illustrate that the target match signals found in PRH are well described as arising from computations implemented in PRH that act on inputs arriving from IT. Somewhat counterintuitively, these signals evolve dynamically within PRH but can be accounted for by instantaneous PRH computation. This is because the inputs from IT change their content over time, and thus the biophysical parameters (e.g. synaptic weights) that are optimal for extracting

diagonal signals at a time that maximizes information content (i.e. 135 ms following stimulus onset) fail to generalize to earlier times where the IT representation differs (e.g. 75 ms).

If changes in the IT target match representation over time were simply due to rescaling (i.e. gradual increases in signal modulation magnitudes produced by stimulus-evoked responses), our model of PRH would fail to reproduce the dynamics that we observe in our PRH data (Fig 9b). Rather, we find that our recorded IT responses reflect multiple types of non-stationarities that combine to produce dynamic computation in PRH (Fig 9c-f). The types of non-stationarities we describe are not exotic. Specifically, the "modulation" non-stationarity that we describe arises in large part from a diversity of latencies across the IT population, and this type of latency diversity has been documented in many different visual brain areas (Schmolesky, Wang et al. 1998). Similarly, neurons that do not simply rescale their response selectivity as a function of time following stimulus onset are also well-documented, particularly in IT (Eskandar, Richmond et al. 1992, Chelazzi, Miller et al. 1993, Sugase, Yamane et al. 1999). What our results demonstrate is that these commonly-observed response dynamics can produce seemingly dynamic computation downstream; or conversely, that observing the dynamic evolution of a signal at one stage of processing should not immediately be attributed to delays in the mechanisms used to compute it (e.g. its implementation in complex, recurrent circuits).

To establish our main effect – a delay between the arrival of "total" versus "linearly separable" information in PRH – we compare the performances of two types of read-out rules applied to the data collected from PRH (i.e. an ideal observer and an SVM

linear classifier, Fig 3a-c). We also compare the performances of linear and a nonlinear classifier with matched structure and number of parameters (Fig 3d) and these two classifiers are very similar to our model of PRH computation (i.e. here we envision computation in PRH as a nonlinear "read-out" of IT, Fig 6). This classification scheme – which is analogous to the first two terms of a polynomial expansion of the optimal classifier boundary - is related to others that have previously been proposed but it does not directly correspond to any that we are aware of. In particular, while other classification methods (e.g. quadratic discriminant analysis and quadratic kernel SVM) also rely on covariance differences to compute a decision boundary, they are not explicitly formulated in terms of a linear-nonlinear (LN) cascade of operations, whereas our method specifies an intermediate population of biologically plausible LN units, upon which a linear read-out could be applied. Notably, when we apply classifiers to the data collected from PRH (Fig 3), we apply them in a manner that might be regarded as a "dynamic" read-out (i.e. we allow the parameters to vary between timebins) whereas when we use these classifiers as models of PRH computation (Fig 6), we enforce that the read-out be "static" (i.e. we fit the parameters at a specific timeslice). Our rationale behind this is that we were interested in evaluating the hypothesis that signals in PRH could be described as arising from a static computation and thus we began by quantifying signals in PRH in the absence of making this assumption and then we then compared these results with a model of PRH when this assumption was enforced. Stated differently, here we first present the classifier analyses merely as quantification tools (Fig 3) and then we proceed to evaluate one as a model of PRH computation (Fig 6).

It is also worth noting that our model is a "functional" and similar to other functional models (e.g. Adelson and Bergen 1985, Rust, Mante et al. 2006), it is designed to capture neural computation in an interpretable manner. To clearly describe how PRH might "compute" linearly separable information arriving from IT, our model separates those signals from the ones that are "inherited" from IT by parsing them into different model PRH neurons (Fig 6a). We note that it is highly unlikely that the brain separates signals in the same way. Rather, the responses of actual PRH neurons likely reflect a combination of both "inherited" and "computed" linearly separable target match signals using mixtures of the mechanisms used in our model for different "classes" of neurons.

In developing our model of PRH, we assumed that the task-relevant connections between IT and PRH were learned and we used this assumption to guide our selections of the spike count window width (25 ms) and its placement (135 ms following stimulus onset). How reasonable are these assumptions? While little is known about the specific mechanisms that regulate synaptic plasticity in PRH during complex cognitive tasks, neural plasticity during reinforcement learning is thought to largely be regulated by dopaminergic inputs (reviewed by Schultz 2007), and we know that PRH contains high densities of both dopamine carrying fibers and dopamine receptors (reviewed by Richmond 2006). Consistent with a specific training window, some have hypothesized that a phasic dopamine response could serve to "switch on" learning at a precise time following stimulus onset (Redgrave and Gurney 2006, Redgrave, Gurney et al. 2008). Thus while much remains to be discovered about synaptic plasticity in PRH, our assumptions are consistent with our current understanding of those mechanisms.

In agreement with earlier reports (Eskandar, Richmond et al. 1992, Chelazzi, Miller et al. 1993, Chelazzi, Duncan et al. 1998), our results suggest that during visual target search tasks, the IT representation is non-stationary; how do these non-stationarities arise in IT? Possibly from multiple sources. First, visual non-stationarities have been reported previously in IT under conditions of passive viewing (Sugase, Yamane et al. 1999), suggesting that "cognitive" (i.e. working memory) signals are not the only contributors. Second, IT non-stationarities may be produced via the mechanisms that combine visual and working memory information within or before IT in the ventral visual pathway. A series of studies documented non-stationarities within V4, IT and PRH as monkeys performed a target search task in which they had to find targets among sets of multiple stimuli (Chelazzi, Miller et al. 1993, Chelazzi, Duncan et al. 1998, Chelazzi, Miller et al. 2001). The authors proposed that target-specific working memory signals may exert their influence via a top-down bias to IT (and/or V4) neurons, followed by competitive interactions within IT that enhance the responses to target stimuli and suppress the response to distractors (Desimone and Duncan 1995). Finally, cognitive non-stationarities may be "inherited" from prefrontal cortex where the persistent, working memory representations of target identity are thought to be housed (e.g. Miller, Erickson et al. 1996) but individual prefrontal neurons are reported to respond only transiently during some fraction of the memory period (Brody, Hernandez et al. 2003, Machens, Romo et al. 2010). As our results demonstrate, regardless of their source, non-stationarities in IT have important consequences for downstream computation.

Finally, we note that our report includes a number of methodological advancements in data analysis and model fitting that may be useful for other studies. First, we apply a method to quantify the amounts of different types of task-relevant

signals contained within heterogeneous and difficult to understand brain areas like IT and PRH (Figs 4-5, elaborated in Pagan and Rust 2014). In our study, this provided an important constraint for our model PRH (Fig 6) and allowed us to quantify multiple types of non-stationarities in IT (Figs 8-9). Second, we introduce derivations that connect these single-neuron measures with population-based analyses (Fig 4c). This allowed us to determine the underlying neural signal dynamics that gave rise to dynamics in the population-based classifier performance measures (e.g. Fig 4b-c). Finally, we introduce a means of "leap-frogging" over a considerable amount of neural processing that we do not understand to determine the computations performed in a high-level brain area (i.e. determining computation in PRH in the absence of a model of processing up to and including IT). We achieved this by fitting an LN model to our recorded IT responses to produce a model PRH that we compared to our PRH data (Fig 6). While our previous attempts at fitting such models were constrained to brute force searches of simple (pairwise) LN combinations, here we used an insight from our previous work (Pagan, L.S. et al. 2013) to fit a more realistic model in which larger numbers of IT neurons combine to form the responses of neurons in PRH.

CHAPTER 4: Quantifying the signals contained in heterogeneous neural responses and determining their relationships with task performance

Marino Pagan and Nicole C. Rust (2014). Journal of Neurophysiology **112**(6): 1584-1598

Abstract

The responses of high-level neurons tend to be mixtures of many different types of signals. While this diversity is thought to allow for flexible neural processing, it presents a challenge for understanding how neural responses relate to task performance and to neural computation. To address these challenges, we have developed a new method to parse the responses of individual neurons into weighted sums of intuitive signal components. Our method computes the weights by projecting a neuron's responses onto a pre-defined orthonormal basis. Once determined, these weights can be combined into measures of signal modulation, however, in their raw form, these signal modulation measures are biased by noise. Here we introduce and evaluate two methods for correcting this bias, and we report that an analytically derived approach produces performance that is robust and superior to a bootstrap procedure. Using neural data recorded from IT and perirhinal cortex as monkeys performed a delayed-match-to-sample target search task, we demonstrate how the method can be used to quantify the amounts of task-relevant signals in heterogeneous neural populations. We also demonstrate how these intuitive quantifications of signal modulation can be related to single-neuron measures of task performance (d').

Introduction

The responses of neurons at higher stages of neural processing in the brain tend to reflect heterogeneous mixtures of many different types of task-relevant signals (e.g. Miller and Desimone 1994, Brody, Hernandez et al. 2003, Buckley, Mansouri et al. 2009, Bennur and Gold 2011, Rigotti, Barak et al. 2013). This diversity is thought to be advantageous insofar as a population that contains a diversity of neural responses is capable of performing a diversity of tasks (Rigotti, Barak et al. 2013). However, response heterogeneity also makes these high-level brain areas difficult to understand using classical single-neuron approaches, which inherently rely on identifying regularities in the response properties of individual neurons across a population (e.g. discovering that the majority of V1 neurons are tuned for orientation).

Here we present a method to deconstruct the responses of heterogeneous neurons as weighted sums of intuitive signals. Our method is useful when applied to experimental designs that involve changing multiple experimental parameters, which is of course a prerequisite for uncovering signal "mixtures". Examples include tasks that require finding a "match" to a target, which involves changing the identities of the "stimuli" and the "target" (e.g. Maunsell, Sclar et al. 1991, Miller and Desimone 1994, Pagan, L.S. et al. 2013). Likewise, tasks that require flexible rule-based mappings of sensory stimuli onto behavioral responses involve manipulating the sensory stimulus and the rule (e.g. Mansouri, Buckley et al. 2007, Bennur and Gold 2011). A slightly less obvious example is a task that requires a subject to remember the specific sequence

with which objects appear; the different conditions in such a task can be envisioned as combinations of object identity and time (Naya and Suzuki 2011).

To address the challenges associated with understanding how the responses of a heterogeneous neural population reflects different task-relevant components, we have developed a method to parse the responses of individual neurons into weighted sums of intuitive components. Our method computes the weights by projecting a neuron's responses onto a pre-defined orthonormal basis. Once determined, these weights can then be combined to quantify different types of signal modulation in a manner that does not depend on sign (e.g. firing rate increases or decreases). From a neural coding perspective, both firing rate increases and decreases convey information and thus unsigned modulation measures more accurately reflect signal magnitude. Additionally, because firing rate increases and decreases tend to be balanced in many high-level brain areas (e.g. Maunsell, Sclar et al. 1991, Miller and Desimone 1994, Romo, Brody et al. 1999), the "average" signed modulation across a population is not a useful quantity (i.e. because it takes on a value near zero) whereas the "average" unsigned (absolute valued or squared) modulation is meaningful.

As we describe in detail below, our method is related to other approaches, including the analysis-of-variance (ANOVA), the multiple linear regression (MLR), the principal components analysis (PCA) and a recent PCA extension (de-mixed PCA; Machens 2010). While these methods have advantages over our method for some applications, one advantage of our method over the others is that it produces unsigned and unbiased estimates of signal modulation magnitudes. Unbiased signal estimates are important when one wants to compare signals across brain areas, across different points

in time, or across different types of signals. However, we note that our method is not ideal for describing exactly "how" neurons are tuned for a particular parameter (e.g. for describing tuning curves).

In addition to introducing a new way to measure neural signals, we demonstrate how these measures can be related to task performance. Quantifying task performance for individual neurons by performing a Receiver Operating Characteristic (ROC) analysis or by calculating the related discriminability measure d' is a common way to compare neural signals – between different brain areas, between different points in time within the same brain area, or with behavior (e.g. Newsome, Britten et al. 1989, Bennur and Gold 2011, Liebe, Logothetis et al. 2011, Adret, Meliza et al. 2012, Gu, Deangelis et al. 2012, Swaminathan and Freedman 2012). Understanding the underlying sources of neural task performance differences (e.g. overall firing rate changes versus changes in different types of tuning modulation) is crucial for accurate interpretation of what these differences mean for neural coding. Here we show how our method can be used to derive a precise understanding of how task performance depends on different types of signal modulation.

Methods

The data we use to describe our method has been reported previously (Pagan, L.S. et al. 2013). All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee. Briefly, we recorded neural responses in IT and PRH as monkeys performed a delayed-match-to-

sample, sequential target search task that required treating the same images as targets and as distractors on different trials (Fig 1a). Monkeys initiated a trial by fixating a small dot and after a short delay, a cue indicating the target for that trial was presented, followed by a random number (0-3) of distractors, and then the target match. Monkeys indicated the presence of the target match by making a saccade to a specific location on the screen before the onset of the next stimulus and were rewarded for correct responses. Altogether, four images were presented in all possible combinations as a visual stimulus ("looking at"), and as a target ("looking for"), resulting in a four-by-four matrix and at least 20 repeated trials of each condition were collected (Fig 1b).

Most of our methods are described in the Results section. Here we describe the statistical procedures we used to evaluate the statistical significance of the observed differences in the mean values of various indices between IT and PRH (Fig 5). Because many of these measures were not normally distributed, we calculated these p-values via a bootstrap procedure. On each iteration of the bootstrap, we randomly sampled the true values from each population, with replacement, and we computed the difference between the means of the two newly created populations. We computed the p-value as the fraction of 1000 iterations on which the difference was flipped in sign relative to the actual difference between the means of the full dataset (e.g. if the mean for PRH was larger than the mean for IT, the fraction of bootstrap iterations in which the IT mean was larger than the PRH mean; Efron and Tibshirani 1994).

Results

The methods we describe here are useful for analyzing the neural data from experiments in which experimental conditions are combinations of multiple stimulus parameters (e.g. sensory stimuli combined with different task instructions). Additionally, they can be applied to both parametric variation (e.g. systematic changes in motion direction) as well as non-parametric variation (e.g. changes in object identity where the relationships between different identities are not well-defined). The ultimate goal of our method is to measure the magnitude by which a neuron's responses are modulated by different experimental parameters and below we refer these modulation magnitudes as "signals". Our method involves parsing a neuron's firing responses to N different combinations of the stimulus parameters (i.e. experimental conditions), which we refer to as a "response matrix", into a weighted sum of N intuitively defined signals. This process begins by constructing an orthonormal basis of N vectors. "Ortho" refers to the fact that the vectors are "orthogonal", and this allows the original matrix to be deconstructed into a weighted sum (i.e. none of the neural responses are counted twice). "Norm" refers to the fact that all the vectors have the same length (i.e. the "norm" of each vector, computed as the square root of the summed squared values, is equal to 1). "Basis" refers to the fact that together, the vectors capture all possible types of response modulation that could occur given the specific experimental design. As described in more detail below, once the orthonormal basis is determined, the weights are calculated for each neuron by taking the projection (i.e. the dot product) of the neuron's average firing rate responses and each basis vector and the "signals" are determined by combining weights of the same type.

Constructing an orthonormal basis

To construct the basis, we begin by constructing a set of N vectors that capture the types of modulation we are interested in. Next we apply the Gram-Schmidt process to convert the set of vectors into an orthonormal basis. To describe the method, we apply this procedure to an example experimental design taken from our previous work: a delayed-match-to-sample (DMS) target search task (Pagan, L.S. et al. 2013). In these experiments, monkeys viewed a series of sequentially presented images and indicated when a "target match" appeared within a sequence of "distractors" (Fig 1a). Altogether, monkeys viewed each of four visual images in the context of each image as a target, resulting in a four-by-four matrix of experimental conditions (Fig 1b). In this matrix, target matches fall along the diagonal and distractors fall off the diagonal. This "response matrix" ***R*** is computed as the average spike count response across 20 repeated trials for each of the 16 experimental conditions. Below, we treat ***R*** as a 16-entry vector to perform our calculations.

To design an orthonormal basis for this task, we began by constructing a first vector which corresponds the grand mean spike count response across all conditions; all entries in this vector take on the same, constant value (e.g. 1/16; Fig 1c). The remaining vectors are designed to capture the types of modulation that neural responses might reflect, which follow from the task design. In the case of our experiment, this included three vectors to describe the visual modulation, reflected by columns in the response matrix (Fig 1c). Notably, while there are four different visual images, only three are

required to capture the visual modulation once the mean firing rate response has also been defined (i.e. degrees of freedom for the visual conditions = 4 – 1). The second type of modulation is reflected by rows in this matrix, and corresponds to response modulations that can be attributed to changing the identity of the target; because target identity must be held in working memory during this task, we refer to this as "working memory" modulation. The third type of modulation differentiates whether a condition was a target match or a distractor, and this corresponds to modulation along the diagonal. The final type of modulation is that which is required to describe responses that are "peppered" across the matrix, such as differential responses to the same visual image under two different distractor conditions, and we refer to this modulation as "residual". More technically, residual modulations reflect all nonlinear combinations of visual and working memory signals that are not diagonal.

Once this initial set of vectors is defined, we apply the Gram-Schmidt procedure to convert it into an orthonormal basis. Specifically, we define each of the N original vectors as $\boldsymbol{v}_i$, and each of the vectors of the resulting orthonormal basis as $\boldsymbol{b}_i$. The Gram-Schmidt process is applied iteratively to each initially defined vector, and consists of two stages: first the vector is orthogonalized relative to all the vectors already incorporated into the final, orthonormal basis, and second, the resulting vector is normalized by its norm $||\boldsymbol{b}_i||$:

$$\boldsymbol{b}_i = \boldsymbol{v}_i - \left(\boldsymbol{v}_i^T \cdot \boldsymbol{b}_1\right) \cdot \boldsymbol{b}_1 - \left(\boldsymbol{v}_i^T \cdot \boldsymbol{b}_2\right) \cdot \boldsymbol{b}_2 - \cdots - \left(\boldsymbol{v}_i^T \cdot \boldsymbol{b}_{i-1}\right) \cdot \boldsymbol{b}_{i-1} \tag{1}$$

$$\boldsymbol{b}_i = \frac{\boldsymbol{b}_i}{||\boldsymbol{b}_i||} \quad ; \quad ||\boldsymbol{b}_i|| = \sqrt{\sum_j b_{ij}^2} \tag{2}$$

where b_{ij} indicates the j-th element of the i-th vector $\boldsymbol{b}_i$.

The final orthonormal basis obtained for our experiment is shown in Figure 1d. A crucial requirement is that the originally defined vectors $\boldsymbol{v}_1 \ldots \boldsymbol{v}_N$ span the full space; if this is not the case, the Gram-Schmidt process will fail to produce a valid orthonormal basis. It is possible to verify this simply by measuring the rank of the matrix obtained by juxtaposing the original vectors $\left[\boldsymbol{v}_1 \cdots \boldsymbol{v}_N\right]$ and checking that it is equal to N.

There is no unique way to parse a set of N vectors into an orthonormal basis. For example, one might consider the "standard basis" as the set of vectors that define each experimental condition (e.g. 10000, 01000, 00100, etc). While this basis is orthonormal, it is not very useful because a projection of a neuron's responses $\boldsymbol{R}$ onto this basis would simply return the mean firing rate response to each experimental condition (i.e. each entry in $\boldsymbol{R}$). Decisions about how to create the initial vectors when designing the basis depend on what one is trying to achieve. We often find it useful to begin by considering the task "inputs" and whether the task "output" (i.e. the solution) can be expressed as a linear or nonlinear combination of the inputs because this approach formalizes the mapping between the computational goals of the task and the neural signals. For the DMS task described above, the task inputs include "visual" and "working memory" signals (i.e. the monkey is presented with the identity of the target, which he holds in working memory, and the identity of the visual image). These are equivalent to the "linear terms" of a two-factor ANOVA analysis. The solution for this task – differentiating whether each condition is a target match or a distractor (i.e. the diagonal matrix) – cannot be expressed by any linear combination of inputs but instead requires a nonlinear computation. However, it is only one of many possible nonlinear

vectors and it is thus essential to parse it from the "residual" vectors, which also reflect nonlinear combinations of visual and memory signals. We note that "diagonal" and "residual" signals would be combined into a single "nonlinear interaction term" in a two-factor ANOVA (for a more extensive description of the relationship between the orthonormal basis and the ANOVA, see the Discussion).

Not all experimental designs allow for orthogonalization, or equivalently, not all experimental parameters can be orthogonalized. For example, Figure 1e depicts a modified experimental design in which some visual images are always presented as distractors and never as targets. In this case, there is no way to produce a component that captures "target match" signals (e.g. one that reflects tuning for whether a condition is a target match or a distractor) that can be orthogonalized with the "visual" components. This is because the experimental design introduces a correlation between image identity and whether the condition is a target match: once you know that the identity of an image is 5, 6, 7 or 8, you know with certainty that the image is a distractor. Stated differently, in this experimental design "target match" and "visual" signals are confounded. Thus an additional advantage of our method is that it introduces a means to evaluate and improve a candidate experimental design through the attempted construction of a useful orthonormal basis.

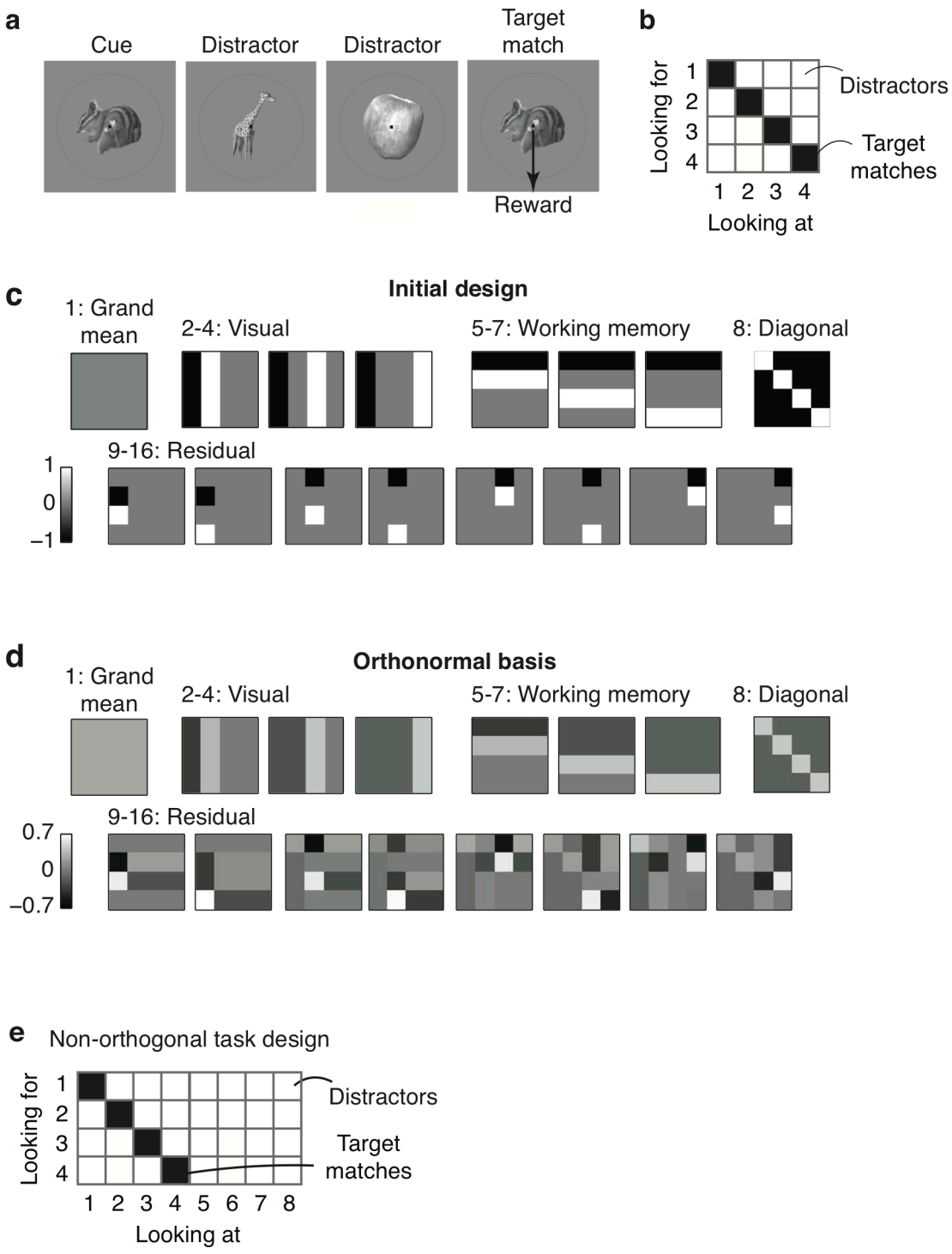

Figure 4-1. *Constructing an orthonormal basis for a delayed-match-to-sample (DMS) task.* a) Each trial of the DMS task began with the presentation of a cue indicating the target for that trial, followed by the presentation of 0-3 distractors, and then the target

match. Images were presented for 400 ms followed by a 400 ms blank. Monkeys were required to maintain fixation throughout the distractors and saccade to a response dot after the target match appeared (within 800 ms) to receive a reward. b) The experimental design included four images each presented as a visual stimulus ("looking at") in the context of every other image as a target ("looking for"), thus defining a 4x4 matrix. In this matrix, target matches fall along the diagonal and distractors fall off the diagonal. c) The matrices produced by the first stage of the orthonormal basis design process (see Text). d) The matrices produced by applying the Gram Schmidt process to the matrices described in panel c. e) An experimental design in which the "target match" and "visual" conditions cannot be orthogonalized (see Text).

Computing and interpreting signal modulation magnitudes

Once the orthonormal basis has been defined, we can compute the corresponding signal modulation magnitudes. A neuron's response matrix $\boldsymbol{R}$ can always be decomposed into a weighted sum of the orthonormal components:

$$\boldsymbol{R} = \sum_{i=1}^{16} w_i \cdot \boldsymbol{b}_i \tag{3}$$

where $\boldsymbol{b}_i$ indicates the i-th component, and w_i indicates the weight associated with the i-th component. The weights w_i are thus determined by computing the projection (i.e. the dot product) of the vector $\boldsymbol{R}$ and each basis component $\boldsymbol{b}_i$:

$$w_i = \boldsymbol{R} \cdot \boldsymbol{b}_i^T \tag{4}$$

Ultimately, we are interested in quantifying how much of a neuron's firing rate modulation can be attributed to changes in specific type of experimental manipulation (e.g. the

amount of firing rate modulation that can be attributed to changes in the visual stimulus), and thus we need to partition the total variance into pre-chosen subspaces (e.g. grouping together the three visual weights). When doing so, it is important to consider that these weights can be negative as well as positive. Positive weights correspond to neural responses that directly resemble an orthonormal basis component whereas negative weights correspond to neural responses that are simply flipped in sign and thus they also reflect relevant firing rate modulations. To convert a set of weights into a measure of a particular type of modulation, we square the weights, sum across the set, and then take the square root. For the DMS task:

$$M_{vis} = \sqrt{\sum_{i \in vis} w_i^2} \quad ; \quad M_{wm} = \sqrt{\sum_{i \in wm} w_i^2} \quad ; \quad M_{diag} = \left| w_{diag} \right| \quad ; \quad M_{residual} = \sqrt{\sum_{i \in residual} w_i^2} \qquad (5)$$

where M_{vis} is the amount of visual modulation, M_{wm} is working memory modulation, M_{diag} is diagonal modulation, and $M_{residual}$ is residual modulation. When computed this way, each type of signal modulation measures the standard deviation (i.e. the spread) of the responses, averaged across repeated trials, and has units of spike count. For example, a “visual signal equal to 2” means that the trial-averaged spike count was spread two standard deviations around the grand mean firing rate as a result of changes in the visual stimulus.

Next we introduce three different ways to normalize these signal modulation magnitudes, each designed to highlight a different aspect of signal modulation. First, one might wish to produce signal modulation measures that are not “raw” (Equation 5) but

instead are compared to the amount of noise. This type of "signal-to-noise" modulation measure can be obtained by simply normalizing by the average trial-by-trial variability of a neuron:

$$M'_{vis} = M_{vis} / \bar{\sigma}_{noise} \;;\; M'_{wm} = M_{wm} / \bar{\sigma}_{noise} \;;\; M'_{diag} = M_{diag} / \bar{\sigma}_{noise} \;;\; M'_{residual} = M_{residual} / \bar{\sigma}_{noise} \tag{6}$$

where $\bar{\sigma}_{noise}$ is computed as:

$$\bar{\sigma}_{noise} = \sqrt{\frac{1}{16} \cdot \sum_{i=1}^{16} \sigma^2_{i,noise}} \tag{7}$$

and $\sigma^2_{i,noise}$ indicates the trial-by-trial variability (variance) associated with the i-th condition. In this formulation, modulations are unitless and they measure the ratio between the signal and noise modulations. For example, a "visual signal equal to 2" now means that changes in the visual signal produced a spread in the trial average spike counts with a standard deviation two-fold larger than the standard deviation of the noise. To anticipate and prevent confusion, we note that the issue of whether a signal modulation estimate is biased by noise is distinct from the issue of normalizing the size of the signal relative to the size of the noise; the former is related to the issue of getting a accurate estimate of signal size (discussed below) whereas the latter informs how much a given amount of signal will be actually "useful" at conveying information. In other words, a fixed amount of signal can provide perfect information in the absence of noise, or it can be almost impossible to detect within a very large amount of noise.

As a second consideration, we note that in some situations, including the DMS task, different types of signals have different numbers of components and it may be

desirable to normalize by the number of components to arrive at a measure of modulation "per degree of freedom":

$$M_{vis} = \sqrt{\frac{1}{N_{vis}} \cdot \sum_{i \in vis} w_i^2} \quad ; \; M_{wm} = \sqrt{\frac{1}{N_{wm}} \cdot \sum_{i \in wm} w_i^2} \quad ; \; M_{diag} = \left| w_{diag} \right| \quad ; \; M_{res} = \sqrt{\frac{1}{N_{res}} \cdot \sum_{i \in res} w_i^2}$$

(8)

where N_{vis} indicates the number of visual components (=3), N_{wm} indicates the number of working memory components (=3), and N_{res} indicates the number of residual components (=8). For example, a "visual signal equal to 2" now means that each visual component was (on average) responsible for spreading the trial-averaged spike count two standard deviations around the grand mean. This normalization can also be combined with the noise normalization in Equation 6 to produce a measurement of the signal-to-noise ratio per component.

Finally, one might wish to produce a measure of signal modulation that is affected by changes in the "pattern" of the response matrix but not by an overall rescaling of the firing rates whereas in their raw form, signal modulations (Equation 5) are directly proportional to the overall grand mean firing rate. Scale-invariant modulation measures can be computed by normalizing each type of signal modulation by the grand mean response to produce quantities that we refer to as "signal strengths". This normalization is described in more detail in the section "Relating signal modulations and task performance".

To illustrate an example of signal modulations, Fig 2 shows the result of our method applied to six neurons collected during the DMS task, including three neurons

whose responses reflect relatively pure selectivity for signals of a single type (Fig 2, top), and three neurons whose responses reflect mixtures of different types of signals (Fig 2, bottom). Shown are the response matrices for each neuron (Fig 2, left column), and the top five orthonormal components rescaled by their weights. As described above, weights can be positive or negative and negative weights invert the polarity of the orthonormal component (e.g. compare the diagonal matrices in the 3rd and 6th rows of Fig 2). Also shown are the signal modulations computed as the square root of the summed, squared weights for each type of signal, normalized by the number of components for each signal type (Fig 2, right column; Equation 8). To produce these plots, response matrices were computed by counting spikes in 25 ms windows systematically shifted relative to response onset. Signal modulations are computed by squaring the weights, so that both positive and negative weights contribute equally to measured modulations. Signal modulations thus provide an intuitive quantification of "how much" of a particular type of signal is reflected in the responses of a particular neuron, regardless of the "sign" of that weight (i.e. responses increases or decreases) and regardless of "how" that modulation is distributed across the different components (i.e. tuning). Importantly, computing modulations in this way is biased (Fig 3-4), and this bias must be corrected to get an accurate measure of modulation (as described below).

As highlighted above, an orthonormal basis is not uniquely defined for a given experimental design. This statement also applies to subsets of different types of components – for example, one could define an orthonormal basis with "visual" vectors that are different from those presented in Figure 1d but capture the visual modulations equally well because the three new visual vectors will define the same linear subspace as the original ones. Thus the combined projection of a neuron's response vector onto

the three visual components uniquely captures the amount of modulation that can be attributed to changes in the identity of the visual stimulus even if the specific visual vectors themselves are not uniquely defined. Under what situations is a particular type of signal modulation uniquely defined? In our experiment, the uniquely defined parsing of different signal types follows from the two-dimensional “looking at” / “looking for” matrix structure this task, in which the “visual” and “working memory” conditions are presented in all possible combinations and are thus independent from one another (Fig 1b). Similarly, because the diagonal matrix is a single dimension, it is also uniquely defined. Finally, the “residual” subspace is uniquely defined because it describes everything that remains after the other uniquely defined subspaces have been accounted for. In contrast, if we were to, for example, combine the first visual, the first working memory and the first residual dimension into a measure of signal modulation, we would obtain a subspace that is strictly dependent on the particular choice of basis, i.e. a different orthonormal basis would produce a different linear subspace when the same three components are considered.

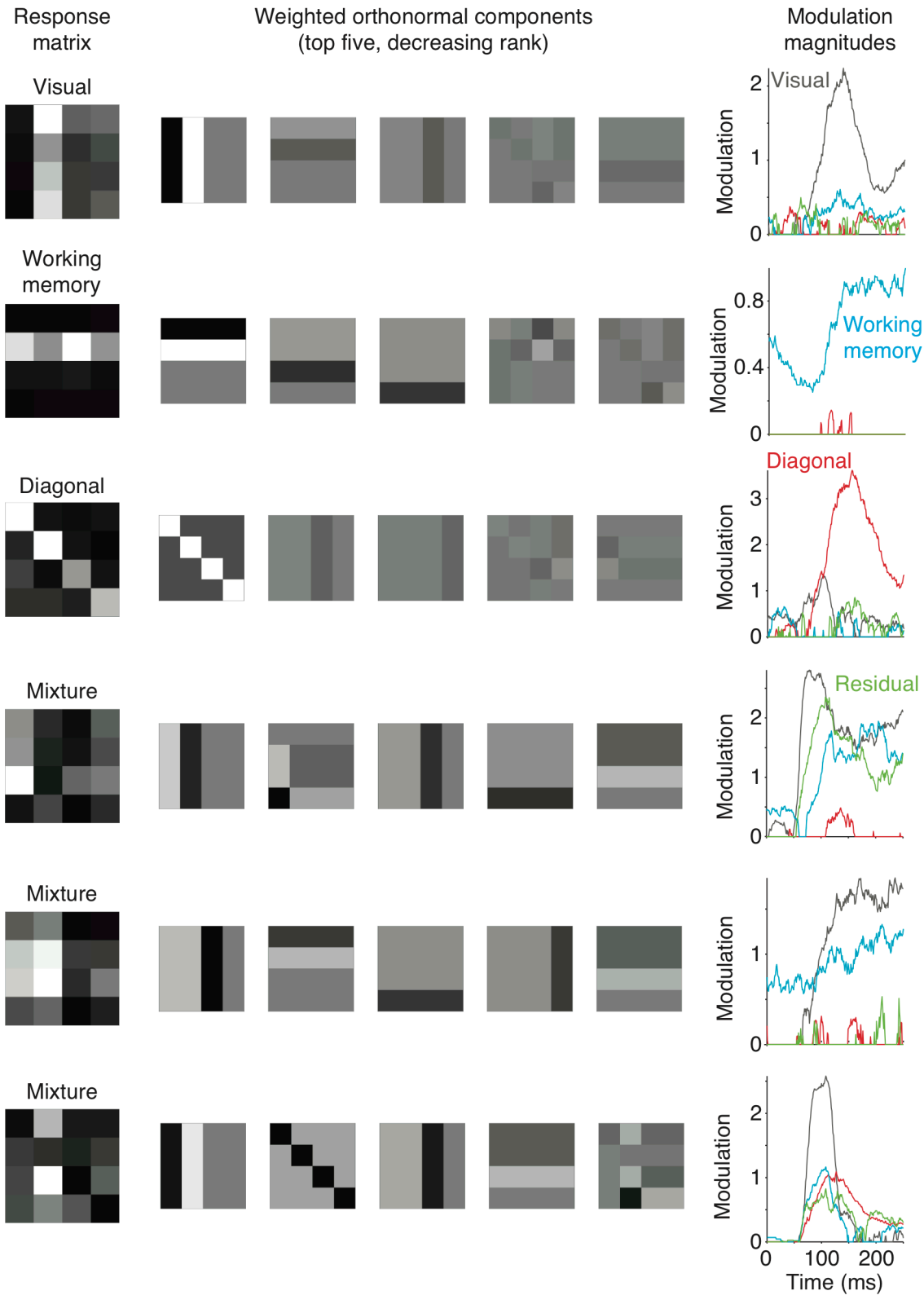
Response matrix
Weighted orthonormal components (top five, decreasing rank)
Modulation magnitudes
Visual
Working memory
Diagonal
Mixture
Mixture
Mixture
Modulation
Visual
Working memory
Diagonal
Residual
0
1
2
0.4
0.8
3
100
200
Time (ms)

Figure 4-2. *Example neurons.* Each row depicts a single example neuron, where the responses of the top three neurons reflect relatively pure selectivity and the responses of the bottom three neurons reflect mixtures of different selectivity types. The first column shows the mean spike count responses computed within a window 50-250 ms following stimulus onset to each of the 16 conditions (the "response matrix"), averaged over 20 repeated trials, normalized to range from the minimum (black) to the maximum (white). The next five columns show the orthonormal components with the five largest weights, plotted as shown in Figure 1d but with intensity scaled by the weight applied to each component. The response matrix can be reconstructed as a weighted sum of these matrices (once the grand mean spike count is also factored in, which is not shown). The rightmost column shows the temporal evolution of the closed-form bias corrected signal modulation magnitudes for each type of signal, computed as the square root of the sum of squares of the bias corrected weights normalized by the number of components for each signal type (Equation 8). To perform this analysis, spikes were counted in 25 ms sliding windows shifted 1 ms for each successive time bin. The example neuron depicted in the fourth row was recorded in IT; the other neurons were recorded in PRH.

Bias and bias correction

When estimating the amount of modulation in a signal, noise and limiting sampling size are known to introduce a positive bias (Panzeri, Senatore et al. 2007). For example, consider a hypothetical neuron that responds with the exact same average firing rate response to each of a set of experimental conditions. Because neurons are noisy, if we were to estimate these mean rates based on a limited number of repeated trials, we would get different values for different conditions and this could lead to the erroneous impression that the neural responses are in fact modulated by the stimuli. Similarly, applying the orthonormal basis method to this data would produce weights

shifted away from zero as a result of noise for at least a subset of the basis vectors. While the mean of the weights themselves would be unbiased (because noise would shift the weights to both more positive and more negative values), the process of converting the weights into signal modulation magnitudes by squaring (Equation 5) would result in a positive mean bias.

To illustrate this bias, Fig 3 includes plots of summed raw (dotted) and bias-corrected (solid) signal modulation magnitudes plotted as a function of time for the IT and PRH populations (using the "closed form" bias correction described below). These results reveal that under physiologically relevant conditions, these biases can be considerable when signals are small or absent (e.g. at stimulus onset, biased estimates of visual modulation are ~1.7 standard deviations in IT and PRH as compared to bias-corrected measures of ~0) and that these biases become smaller when signals are larger (e.g. at the peak of the visual signal, the bias is ~0.25 standard deviations in IT and PRH, which is only ~3% of the bias-corrected value). This is because the bias is additive in the domain of the squared weights but to compute signal modulations, we take the square root. The square root operation has the effect of enhancing the effect of the bias when the modulation is small and shrinking it when the modulation is larger. The reason why we prefer to take the square root rather than operating on the squared modulations is that we find that measures of signal modulations in units of "spike counts" are preferable to units of "squared spike counts" in that they more clearly map onto our intuitive definitions of signals (e.g. signals double when firing rates double).

To estimate bias, we compared two methods: an analytical solution and a bootstrap technique. Under the assumption that trial-by-trial variability is Gaussian

distributed, which is a reasonable approximation of Poisson distributions when the mean spike counts are sufficiently large, the amount of bias can be derived and unbiased measures of the squared weights $\hat{w}_i^2$ can be computed as (see Appendix):

$$Bias_{closed\ form} = \frac{\boldsymbol{R} \cdot \left(\boldsymbol{b}_i^T\right)^2}{T} \quad ; \quad \hat{w}_i^2 = \left(\boldsymbol{R} \cdot \boldsymbol{b}_i^T\right)^2 - \frac{\boldsymbol{R} \cdot \left(\boldsymbol{b}_i^T\right)^2}{T} \tag{9}$$

where T equals the number of repeated trials for each experimental condition.

Because the analytical solution assumes that spike counts are Gaussian distributed whereas spike count distributions are known to deviate from this assumption, particularly at low firing rates, we also introduce a bootstrap procedure. The first step in estimating the bias for a given weight involves resampling with replacement T responses to each condition and recomputing the squared weight for these bootstrapped responses $\tilde{w}_i^2$ using Equation 4. Next, the bias can be estimated by subtracting the modulation computed from the actual responses from the bootstrapped modulation estimates, and finally a corrected estimate for each squared weight $\hat{w}_i^2$ can be computed simply by subtracting the bias (Efron and Tibshirani, 1994):

$$Bias_{bootstrap} = \tilde{w}_i^2 - w_i^2 \quad ; \quad \hat{w}_i^2 = w_i^2 - \left(\tilde{w}_i^2 - w_i^2\right) \tag{10}$$

In practice, we find that the bias estimated on any one resampling can be noisy and thus we find it useful to calculate the bias a number of times (e.g. 100) and average the bias across those calculations.

To test our bias correction procedures, we performed simulations in which we created “ground truth” neurons with known amounts of underlying modulation, simulated

their trial-by-trial variability as Poisson, and compared the ground truth and estimated modulation magnitudes. To test the bias correction in a relevant regime, we performed these simulations by creating a population of 150 "ground truth" neurons that were inspired by actual neurons we recorded in IT and PRH (examples include the neurons shown in Fig 2). Specifically, we computed each simulated neuron's underlying responses by applying a bias correction to 150 randomly selected raw response matrices measured in our experiments, and we then used these mean values to generate N Poisson simulated trials. Figure 4a shows the fractional bias (total bias / total signal), computed for a population of 150 neurons, averaged over 100 simulated experiments. This plot reveals that, as expected, total bias decreases as a function of the number of repeated trials (Fig 4a, black). With only 2 trials, the magnitude of the bias exceeded the magnitude of the signal (fractional bias ~1.5), and fractional bias dropped to ~0.15 for 20 trials and ~0.025 for 100 trials. However, at all numbers of trials, the closed-form bias correction did a very good job at correcting bias (maximal fractional bias remaining after correction = 0.01 for 2 trials; Fig 4a, red). In contrast, fractional bias remained high after the bootstrap correction for small numbers of trials (~0.7 for 2 trials), but converged to the closed-form correction for more than 25 trials (Fig 4a, cyan). Poor bootstrap performance with small sample size is a well-known phenomenon (Chernick 2007).

For a closer look at the closed-form bias correction, Figure 4b displays the distribution of fractional error (total error / total signal) after correction across the 100 simulated experiments when 20 trials for each condition were collected. This distribution is centered around 0, thus confirming that the bias has been successfully removed, and it shows that the remaining average fractional error is small (<0.012 in magnitude) for

individual experiments. These results support the validity of our procedure for estimating signal modulation magnitudes averaged across a population. However, we caution the reader that while the average signal modulation estimates are very accurate, no method can correct for the specific "noise" patterns within the data for a particular neuron. To illustrate this, Figure 4c (left) displays the distribution of fractional error remaining after correction for a representative simulated neuron across the 100 simulated experiments. On average, the error was zero (thus showing no bias), however, on individual simulated experiments, the fractional error ranged from -0.45 to 0.46. Figure 4c (right) shows the "ground truth" response matrix for this neuron as well as one response matrix collected during a simulated experiment. As depicted by the "ground truth" matrix, this neuron was largely visual modulated and responsive to the visual presentation of image 4 but the response matrices collected in this simulated experiment reflects other types of modulation as a result of trial-by-trial variability; even after bias correction, these translate to signal modulation estimates that deviate from the true underlying value. These results demonstrate that the signal modulation magnitudes computed for any individual neuron need to be interpreted cautiously.

The simulations reported above were performed using spike counting windows of 50 ms and with that size counting window, we found that the closed-form bias correction was better at estimating bias than the bootstrap bias correction (Fig 4a). We wondered whether the bootstrap might perform better than the closed-form correction for smaller counting windows where spike count distributions deviate more from the Gaussian assumption (e.g. are Poisson) and thus we compared both types of correction for spike count windows of 2 ms. In these narrow windows, the total fractional bias increased dramatically relative to the broader windows (bias was 16-fold larger than signal for 2

trials; Fig 4d black) but the closed-form bias correction continued to perform well (maximal fractional bias remaining after correction = -0.04 for 2 trials; Fig 4d, red). In contrast, the bootstrap correction performed considerably worse at all numbers of trials, with the most discrepant differences for small numbers of trials; even with 10 trials, the average fractional bias remaining after bootstrap correction was 47% the magnitude of the signal (Fig 4d, cyan). These results suggest that for small spike count windows and the numbers of trials typically collected in these types of experiments (n=5-20), the bootstrap correction is highly inaccurate. In contrast, the closed-form bias correction is highly accurate within this regime despite its assumption of Gaussian distributed trial-by-trial variability.

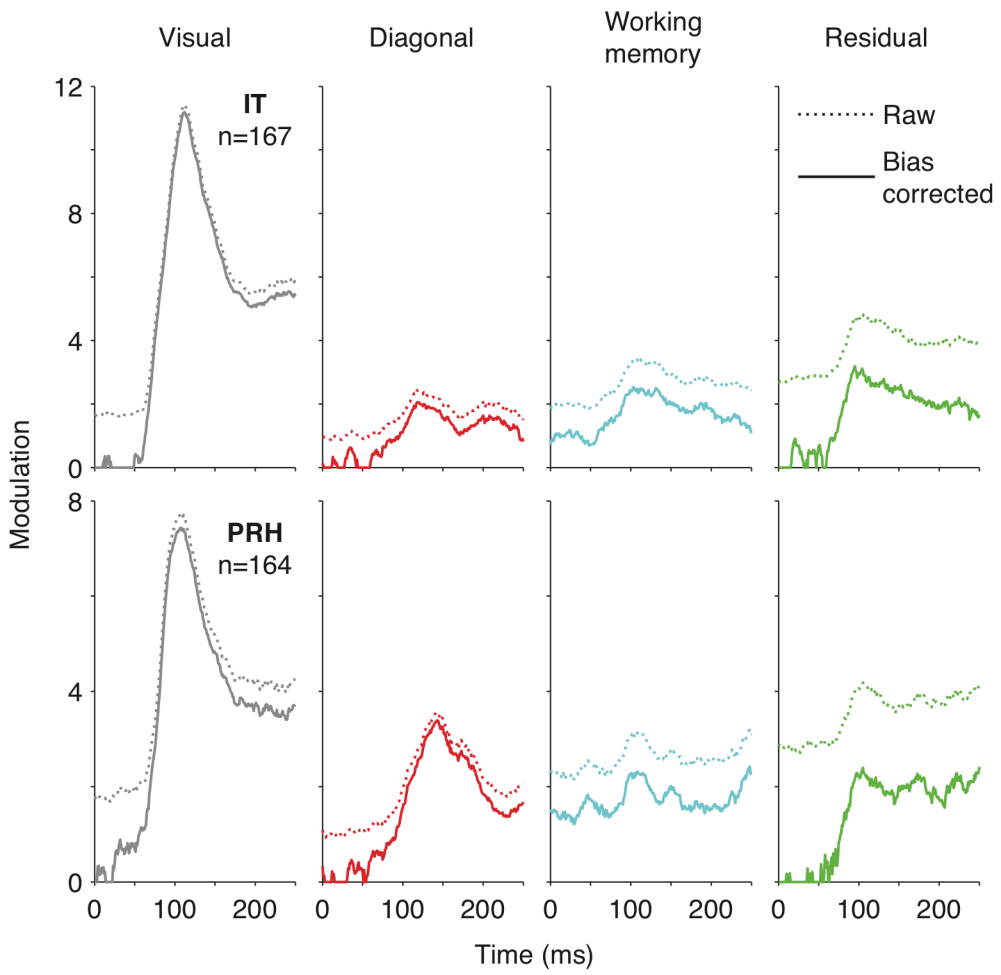

Figure 4-3. *Empirical demonstration of bias.* Raw (dotted) and closed-form bias-corrected (solid) measures of signal modulation summed across the IT (top row) and PRH (bottom row) populations plotted with the same conventions as Fig 2, right.

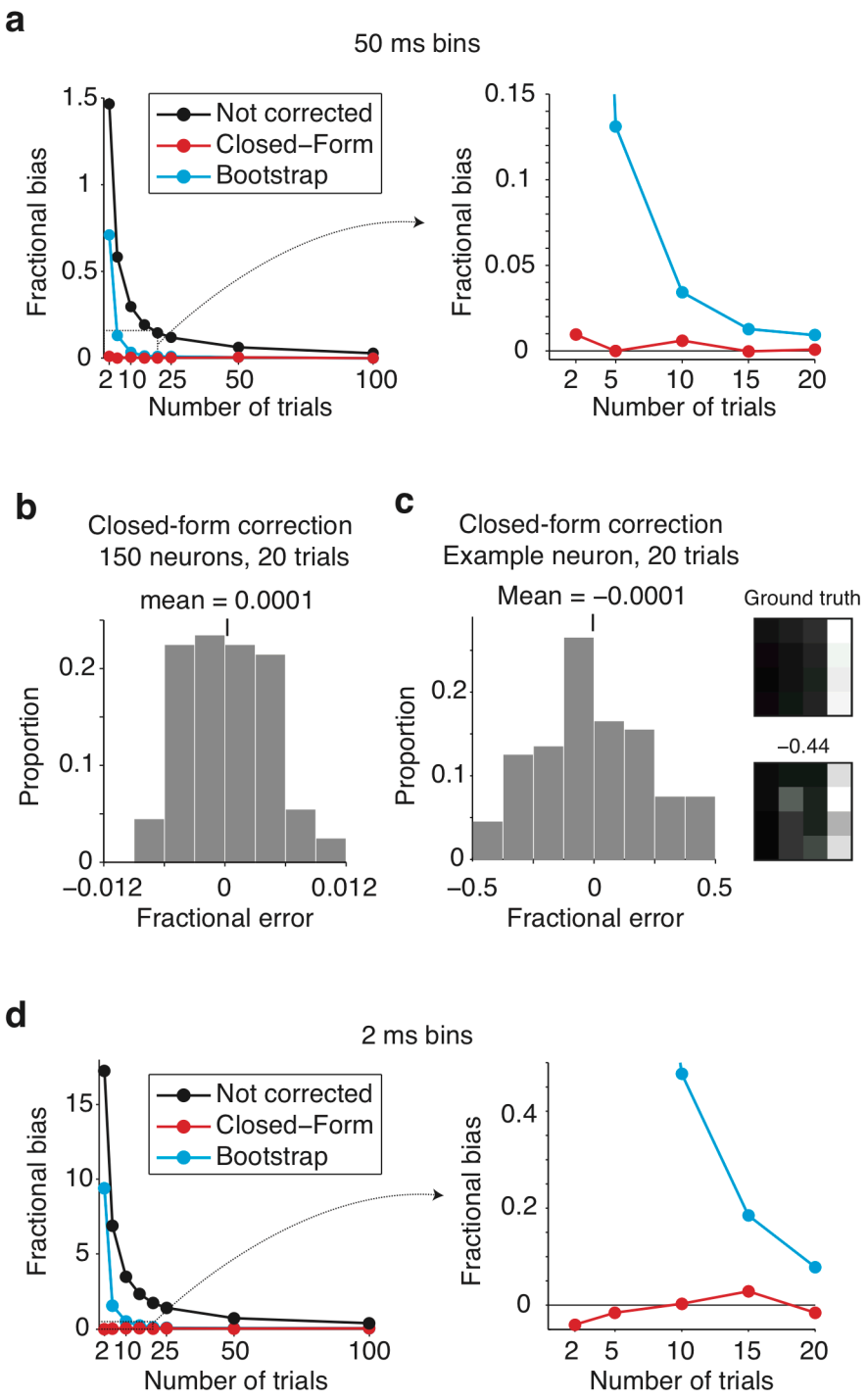

Figure 4-4. *Evaluation of bias-correction procedures.* To evaluate the accuracy of our signal modulation measures, a population of 150 simulated neurons with known amounts of signal modulation were created from measured responses in IT and PRH. a) Fractional bias, calculated as the ratio of the total bias (summed across all signal

modulations) divided by the total signal (summed across all signal modulations), plotted for the uncorrected simulated population (black), the closed-form bias correction (red), and the bootstrap bias correction (cyan) as a function of the number of Poisson trials collected in each simulated experiment when spikes were counted in 50 ms bins centered 125 ms after stimulus onset. Shown are the averages over 100 simulated experiments. The plot on the right shows an enlargement of the dotted region indicated on the left. b) A histogram of the average (across the 150 simulated neurons) fractional error (total error / total signal) remaining after the closed-form correction for each of 100 simulated experiments with 20 Poisson trials to show that the fractional bias measured per experiment is always near zero. c) *Left,* A histogram of the fractional bias remaining for one representative simulated neuron to show that fractional error measured per neuron can be large. *Right,* the "ground truth" response matrix for this neuron plotted along with one example matrix measured from a simulated experiment that produced an extreme fractional error. d) The results of the same analysis presented in panel a, but performed from responses counted in 2 ms windows. As in panel a, the plot on the right shows an enlargement of the dotted region indicated on the left.

Relating signal modulations and task performance

Quantifying the performance of individual neurons on a task by calculating the discriminability measure d' is a commonly used approach to compare neurons within or between brain areas. For tasks that involve multiple experimental parameters or require the combination of multiple information sources to compute a solution, arriving at a quantitative understanding of how different signal types relate to task performance can be challenging. Here we derive this relationship for the DMS task (Fig 1a). We then go on to demonstrate how this type of quantitative understanding can be used to (e.g.)

determine which of many possible accounts can explain why two populations have different average d', by applying the analysis to data collected in IT and PRH.

The delayed-match-to-sample task described in Figure 1a requires a subject to determine whether each test image is a target match or a distractor, and thus can be envisioned as a two-way discrimination between the set of all target matches versus the set of all distractors. Because target matches and distractors correspond to conditions on versus off the diagonal of the response matrix, respectively, we refer to this as "diagonal d'". Diagonal d' is calculated as the absolute value of the difference between the mean response to all target matches and the mean response to all distractors, divided by their pooled standard deviation:

$$|d'| = \frac{|\mu_{Match} - \mu_{Distractor}|}{\sigma_{pooled}} \quad , \quad \text{where} \quad \sigma_{pooled} = \sqrt{\frac{4 \cdot \sigma^2_{Match} + 12 \cdot \sigma^2_{Distractor}}{16}} \tag{11}$$

Because target match modulations in IT and PRH result from both increases and decreases in the firing rates (e.g. Fig 2, 3rd versus 6th rows), the absolute value of diagonal d' best quantifies the linearly discriminable match/distractor information in each neuron. Similar to the signal modulation bias described above, merely taking the absolute value of d' produces a biased estimate of performance in which any modulations, including noise, translate into positive d'. In particular, note that this bias is directly dependent on the numerator of the d' (i.e. the estimated absolute difference can be larger than 0 even if the true difference was 0), while the denominator corresponds to the classic estimator of the standard deviation and it does not have a direct impact on the bias (see Appendix). To correct for the bias of the d' we can thus focus on correcting

for the bias of the numerator, which requires a calculation analogous to the one described above for the case of signal modulations (Equation 9). In particular, it is possible to show (see Appendix) that the bias of the squared numerator is equal to:

$$\frac{\sum_{i=1}^{4}\frac{1}{16}\cdot m_i + \sum_{i=1}^{12}\frac{1}{144}\cdot d_i}{T} \tag{12}$$

where m_i indicates the response to the i-th match and d_i indicates the response to the i-th distractor, and T indicates the number of trials. Therefore, a corrected estimate of the absolute d' can be obtained as:

$$\left|\hat{d}'\right| = \sqrt{\frac{\left(\mu_{Match} - \mu_{Distractor}\right)^2 - \frac{\sum_{i=1}^{4}\frac{1}{16}\cdot m_i + \sum_{i=1}^{12}\frac{1}{144}\cdot d_i}{T}}{\sigma^2_{pooled}}} \tag{13}$$

where $\left|\hat{d}'\right|$ is set to 0 if the numerator takes on a negative value. Below, we also apply the bias-correction to estimate the orthonormal weights (Equation 9).

To derive the relationship between the signal modulations of a neuron (Fig 2) and its diagonal d', we begin by computing the orthonormal basis weights (Equation 4). As described above, rescaling a neuron's firing rate responses (e.g. multiplying a neuron's response matrix by two) will result in a rescaling (i.e. a doubling) of all its weights, and consequently, its signal modulations will also increase. To describe the signal modulations in a manner that does not depend on the overall scaling of firing, we normalized the weights by the grand mean spike count $\overline{SC}$. We then considered the grand mean spike count as a separate term.

Using the orthonormal basis, the diagonal d' of a neuron can be deconstructed as a function of three intuitive "signal strengths" (see the Appendix for the derivation):

$$|d'| = \sqrt{\frac{D}{ND + 1/\overline{\overline{SC}}}} \tag{14}$$

The first signal strength, "D", which we call the "Diagonal strength" (Fig 5, red) is computed as:

$$D = \frac{1}{3} \cdot \left(\frac{w_{diag}}{\overline{\overline{SC}}} \right)^2 \tag{15}$$

where w_{diag} corresponds to the weight applied to the diagonal basis component (Fig 1d). This signal strength determines the distance between the average of the diagonal responses (the target matches), and the average of the off-diagonal responses (the distractors), averaged across all images and all trials (Fig 5b, red line), and this term is proportional to diagonal d' (Fig 5c).

The second signal strength, "ND", which we call the "Non-diagonal strength" (Fig 5, cyan) is computed as:

$$ND = \frac{1}{16} \cdot \sum_{\substack{i \neq diag \\ i \neq mean}} \left(\frac{w_i}{\overline{\overline{SC}}} \right)^2 \tag{16}$$

where the weights used are those corresponding to the visual, working memory and residual components (Fig 1d). This term determines the spread of the firing rate responses within the target matches and within the distractors (Fig 5b; cyan line) and it is inversely related to diagonal d' (Fig 5c).

The final term $1/\overline{SC}$ (Fig 5, lavender) is designed to capture the trial-by-trial variability of a neuron. When trial-by-trial variability is generated by a Poisson process, the grand mean spike count can be used as a good approximation of the variance across trials within each condition, averaged across the 16 conditions and this term can be described by the inverse of the grand mean spike count (see Appendix). This term is also inversely related to diagonal d', as an increase in the spread within each condition will produce an overall increase in the spread across the set of all target matches and the set of all distractors.

We now demonstrate how understanding the relationship between different signal types and single-neuron task performance can be used to gain insight into neural processing by applying these analyses to our data from IT and PRH. We begin with the observation that diagonal d' was significantly higher in PRH as compared to IT (mean IT=0.11, PRH=0.19, $p<0.001$, Fig 5a). We can use the derivation of diagonal d' presented above to discriminate between different possible explanations of why diagonal d' is higher in PRH. Our decomposition suggests three possible factors that might account for this result (that are not mutually exclusive): 1) the diagonal strength (Fig 5c, red) could be higher in PRH than in IT, 2) the non-diagonal strength (Fig 5c, cyan) could be lower in PRH than in IT and/or 3) grand mean firing responses (Fig 5c, lavender) could be higher in PRH than in IT. First and foremost, the diagonal strength was significantly higher in PRH than in IT (Fig 5d), suggesting that this factor contributed to higher average neuron diagonal d' in PRH. Second, the non-diagonal strength was not significantly different between IT and PRH (Fig 5e), suggesting that this factor could not account for the difference in neuron diagonal d'. Finally, the grand mean firing rates were slightly lower in PRH as compared to IT but not significantly so (Fig 5f), and notably,

lower firing rates in PRH are the opposite of what would be required to account for higher average PRH neuron diagonal d' (Fig 5c). Taken together, these results suggest that higher neuron diagonal d' results from a two-fold increase in diagonal structure within the response matrices of PRH neurons as compared to IT neurons, as opposed to alternative explanations (such as increases in firing rate in PRH or more non-diagonal modulation in IT).

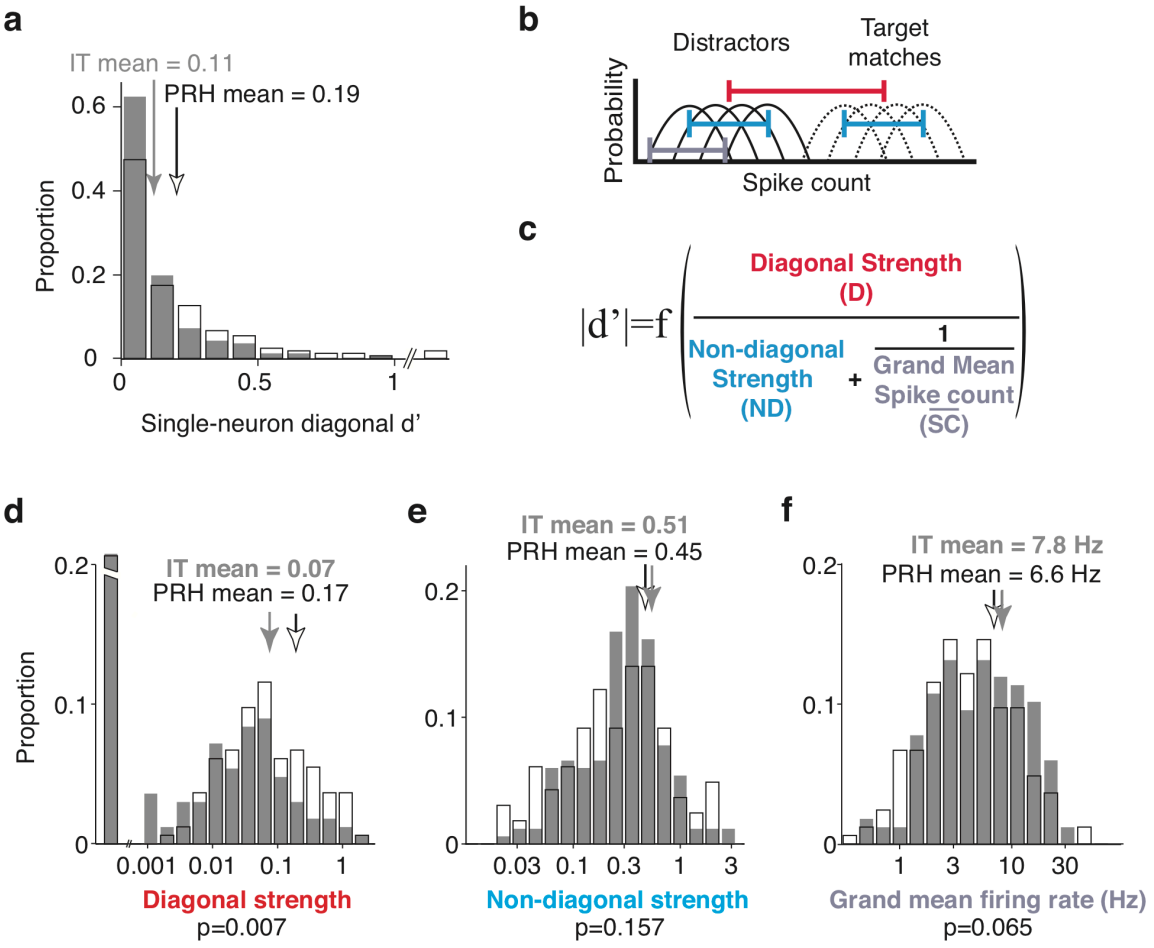

Figure 4-5. *Relating signal modulation magnitudes with task performance (d').* a) Diagonal d', computed as the absolute value of d', followed by bias correction (Equation

13). Histograms are shown for 167 neurons recorded IT and 164 neurons recorded in perirhinal cortex. PRH neurons with neuron diagonal d' of 1.1, 1.3, and 2.1 are included in the last bin. Arrows indicate means. b) The diagonal d' calculation was based on the distributions of responses over trials to the target matches (dashed) and to the distractors (solid lines). c) Diagonal d' can be expressed as a function of three intuitive components. b-c) The diagonal strength (red) is computed as a function of the normalized diagonal weight (Equation 15) and this component determines the distance between the average response to all target matches and the average response to all distractors (red line); diagonal d' is proportional to this component (see Text). The non-diagonal strength (cyan) is computed as a function of the combined normalized non-diagonal weights (Equation 16) and this component determines the spread within the target matches and within the distractors (cyan line); diagonal d' is inversely related to this component (see Text). The final component (lavender) captures the trial-by-trial variability of the neuron (lavender line). Within a Poisson process and after normalization, this term can be estimated by the inverse of the grand mean spike count across all conditions (Equation 40), and thus diagonal d' increases monotonically with the grand mean spike count (see Text). d) Diagonal strength, b) Non-diagonal strength and c) Grand mean firing rate, in IT (gray) and PRH (white). Arrows indicate means. In subpanel a, the first bin includes neurons with a diagonal strength less than 0.001 and the broken axis extends to a proportion of 0.46 in IT and 0.34 in PRH.

Discussion

In our own work, we have found this method of estimating signal modulation magnitudes to be useful for a variety of applications. For example, we have used these signal quantifications as a benchmark to assess model performance (Pagan, L.S. et al. 2013). We have also used these methods to compare the latencies with which specific types of signals arrive in different brain areas to infer the direction of information flow

between them (Pagan, L.S. et al. 2013). As described above, these methods can also be used to uncover the underlying source of differences in single-neuron performance measures between brain areas to gain insights into neural coding. These are but a few examples of the potential uses of this method.

Relationship to other analyses

The method we describe here is similar to a multi-way ANOVA, but it incorporates two important extensions: it parses the signal into more terms and it produces a bias-corrected estimate of signal modulation. For the DMS task described above, a two-way ANOVA would parse the total response variance into two linear terms, a nonlinear interaction term, and an error term. The two ANOVA linear terms map directly onto the summed squared projections onto the visual and working memory orthonormal basis vectors (e.g. in Equation 5, the computation of M_{vis} and M_{wm} before taking the square root). Similarly, the ANOVA nonlinear interaction term maps onto the summed squared projections onto the "diagonal" and "residual" terms in our analysis. We note that parsing the diagonal signals from the other nonlinear terms is crucial in our analysis because this signal reflects the task solution, whereas the other types of nonlinear terms do not. The final term in the ANOVA analysis, the error term, is equal to the square of the $\bar{\sigma}_{noise}$ term described in Equation 6. We remind the reader that in this raw form, the values of the orthonormal basis, as well as the ANOVA, are biased due to trial-by-trial variability (i.e. response matrix structure that arises from noise). The ANOVA deals with this bias by computing the probability (the p-value) that each term is

significantly higher than expected by chance, given the trial-by-trial variability, by considering the ratio between each term and the error term (the "F statistic"), based on the assumption that the noise is Gaussian-distributed. However, the ANOVA does not produce bias-corrected estimates of signal modulation whereas here we describe two ways to estimate and correct for this bias.

Our method also has similarities with an approach related to the ANOVA, multiple linear regression (MLR). Similar to our procedure, MLR seeks to describe a neuron's responses as a weighted sum of multiple terms. In practice, it is most often applied to continuous variables (e.g. motion direction or color), and often in cases in which one has an specific underlying model of how different stimulus parameters combine to determine a neuron's response (e.g. knowledge that neurons have Gaussian shaped tuning functions for motion direction). MLR can also be applied in non-parametric cases and when used in this way, multiple terms are required to capture response modulation of a single variable type (e.g. for object identity, response = baseline + weight_1*identity_1 + weight_2*identity_2 + …) and in fact, our method could be described as an MLR with the regressors specified by an intuitive orthonormal basis. When viewed from this perspective, our method can provide multiple insights into those wishing to perform this type of MLR. First, a crucial consideration with MLR is the degree to which the different regressors are correlated with one another, because the values of the weights (i.e. the "beta coefficients") can be misleading in case of strongly correlated regressors. One solution to this problem is to orthogonalize the variables of interest although we note that for some data sets, the experimental variables simply cannot be orthogonalized (e.g. Fig 1e). Our method provides a straightforward way to evaluate the degree to which different candidate experimental designs can be orthogonalized for MLR. Second,

determining the weights for a complete orthonormal basis guarantees a full account of a neuron's spike count modulation whereas an MLR against a few (e.g. linear) terms might provide only a partial account. Finally, if one desires to convert MLR "beta coefficients" into positive-valued measures of modulation, these measures will be biased in the exact same manner we describe above and here we introduce a way to correct for that bias.

Our method also has similarities with principal components analysis (PCA) and related techniques (Machens 2010). A PCA applied to the mean firing rate responses of a population of neurons to N experimental conditions returns an orthonormal basis of N "eigenvectors" and each neuron's mean firing rate response to the N conditions can be decomposed into a weighted sum of these vectors by projecting the neuron's responses onto the basis, as described above. PCA differs from our method in that the eigenvectors are produced via a procedure that iteratively determines the stimulus dimensions that account for the most response variance across the population with the constraint that each successive vector must be orthogonal to all the others. Consequently, PCA dimensions are not guaranteed to be intuitive. As an illustration, Fig 6a shows the results of a PCA applied to our IT and PRH populations. While the two largest eigenvectors for each population are primarily visual, they are not purely so, and eigenvectors of rank three and lower capture mixtures of different types of modulation. Thus PCA is not very useful in providing an intuitive description of the types of signals reflected in these populations. Rather, PCA is most often used as a "dimensionality reduction" technique. For example, in the case of the reverse correlation method "spike-triggered covariance" (Schwartz, Pillow et al. 2006) one applies a PCA to the spike-triggered stimuli in an attempt to find a small number of stimulus dimensions that can account for individual neuron's responses within a linear-nonlinear model framework.

One extension of the PCA framework, demixed PCA (dPCA; Machens 2010, Brendel, Romo et al. 2011) has recently been introduced as a solution to the "mixing" issues described above for PCA. dPCA allows one to specify the experimental parameters that should not be mixed and thus to perform dimensionality reduction within specific linear subspaces. It is advantageous over our method in scenarios in which (e.g.) one wants to determine whether the responses to a specific type of stimulus parameter can be captured with a simple (i.e. low-dimensional) description, or equivalently to uncover specific types of "tuning". For example, dPCA has provided important insights into how the memory delay period activity of neurons in prefrontal cortex depends on time (Machens, Romo et al. 2010). The results of a dPCA applied to our data in IT and PRH are shown in Fig 6b. The input to dPCA included the neural responses to all conditions as well as the task parameters (i.e. the visual and target identities) associated with each condition. This information, which is not provided to traditional PCA, allows dPCA to search for a set of components that capture most of the modulation while avoiding mixing different types of signals (e.g. visual and working memory). In contrast to a regular PCA (Fig 6a), the first three components in each area are almost exclusively visual, and the fourth component for PRH corresponds to the "diagonal" component of the orthonormal basis. However, one can also see from this analysis that if the desired outcome is a characterization of "how much" of specific, pre-defined signal types are present in a population, the orthonormal basis provides a better approach for two reasons: 1) the components retrieved by dPCA still present some degree of "noise" and thus if the relevant axes are known in advance it is better to measure their modulations directly and 2) in situations in which one wants to make a quantitative comparison between two populations, some compromise has to be

established when different dPCA components are retrieved for each populations (e.g. compare IT and PRH in Fig. 6b).

Finally, a complementary approach for quantifying signals is to measure single neuron performance either by a Receiver Operating Characteristic analysis (e.g. Newsome, Britten et al. 1989, Bennur and Gold 2011, Swaminathan and Freedman 2012) or by the related (boundless) discriminability measure d' (e.g. Liebe, Logothetis et al. 2011, Adret, Meliza et al. 2012, Gu, Deangelis et al. 2012). Under the assumption that trial-by-trial variability is Gaussian distributed, one can convert between the two measures with a simple nonlinear function (i.e. the complementary error function; Dayan and Abbott 2001). As our results show, in a multi-parameter task like DMS, single-neuron task performance does not necessarily depend on a single type of signal but instead can reflect the combination of multiple signal types. Additionally, it is important to note that if one wishes to compute a measure of task performance that is unsigned (i.e. by taking the absolute value or squaring), these task performance measures will be biased. However, this bias can be estimated and corrected using the approaches we describe here.

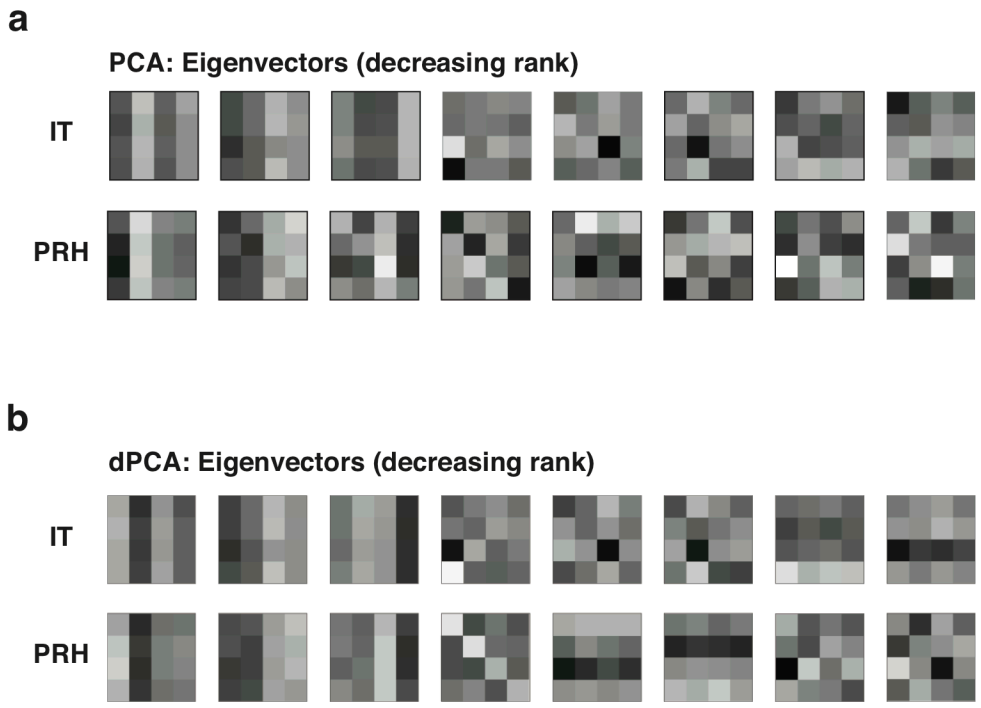

Figure 4-6. *Results of PCA and dPCA.* a) Illustration of the orthonormal components corresponding to the eight largest eigenvectors obtained applying PCA to our IT and PRH data. b) The eight largest orthonormal components resulting from the application of dPCA.

Appendix

Derivation of the bias correction for signal modulations

When estimating the amount of modulation (or information) in a signal, noise and limited sample size are known to introduce a positive bias (e.g. Treves and Panzeri 1995). Here we quantify the magnitude of this bias for the weights associated to the

different signal components (Equation 4), which are used to produce the estimated modulation components (Equation 5) of a single neuron.

We begin by making the simplifying assumption that responses to each condition j are normally distributed with mean μ_j and variance σ_j^2, and that for each condition, μ_j is approximately equal to σ_j^2. We indicate the estimate of the mean response μ_j to each condition as r_j defined as the average of the responses sampled over T trials. The value of the estimate r_j will itself be normally distributed, with mean equal to the true mean μ_j and variance equal to σ_j^2 / T.

By expanding Equation 4, the estimated weight associated to each component *i* can also be written as:

$$w_i = \boldsymbol{R} \cdot \boldsymbol{b}_i^T = \sum_{j=1}^{16} r_j \cdot b_{ij} \tag{17}$$

where r_j indicates the neuron's average response to the j-th condition, and b_{ij} indicates the j-th entry of the i-th basis component. Since w_i is a linear combination of normally distributed variables, it will also be normally distributed. The mean of w_i will thus be equal to the linear combination of the means of the estimated r_j (i.e. the true mean responses μ_j) and the entries b_{ij}, while the variance will be equal to the linear combination of the variances of r_j (i.e. σ_j^2 / T) and the squared entries b_{ij}^2:

$$Mean\,(w_i) = \sum_{j=1}^{16} \mu_j \cdot b_{ij} \quad ; \quad Variance(w_i) = \frac{\sum_{j=1}^{16} \sigma_j^2 \cdot b_{ij}^2}{T} \tag{18}$$

Note that $\sum_{j=1}^{16} \mu_j \cdot b_{ij}$ is the value of the "true" weight, and this estimate of w_i is unbiased. However, a bias is introduced by squaring the estimated weight w_i. The square of a normally distributed variable with non-zero mean takes the form of a non-central chi-squared distribution, whose mean is equal to the sum of the squared mean of the original normally distributed variable plus its variance. In our case:

$$Mean\,(w_i^2) = \left[Mean\,(w_i)\right]^2 + Variance\,(w_i) = \left(\sum_{j=1}^{16} \mu_j \cdot b_{ij}\right)^2 + \frac{\sum_{j=1}^{16} \sigma_j^2 \cdot b_{ij}^2}{T} \tag{19}$$

Under the assumption that the variance for each condition is equal to its mean:

$$Mean\,(w_i^2) = \left(\sum_{j=1}^{16} \mu_j \cdot b_{ij}\right)^2 + \frac{\sum_{j=1}^{16} \sigma_j^2 \cdot b_{ij}^2}{T} = \left(\sum_{j=1}^{16} \mu_j \cdot b_{ij}\right)^2 + \frac{\sum_{j=1}^{16} \mu_j \cdot b_{ij}^2}{T} \tag{20}$$

where the first term corresponds to the "true" squared weight, and the second term represents the additive bias. If we substitute the true mean responses with their estimates, we obtain an estimator of the bias:

$$Bias = \frac{\sum_{j=1}^{16} \mu_j \cdot b_{ij}^2}{T} \qquad Bias\ estimator = \frac{\sum_{j=1}^{16} r_j \cdot b_{ij}^2}{T} = \frac{\boldsymbol{R} \cdot \left(\boldsymbol{b}_i^T\right)^2}{T} \tag{21}$$

where $\boldsymbol{R}$ indicates the neuron's response matrix ("flattened" into a vector), and $\left(\boldsymbol{b}_i^T\right)^2$ indicates the i-th basis function, squared element-by-element. Finally, an unbiased estimator of w_i^2 is given by:

$$\hat{w}_i^2 = \left(\boldsymbol{R}\cdot\boldsymbol{b}_i^T\right)^2 - \frac{\boldsymbol{R}\cdot\left(\boldsymbol{b}_i^T\right)^2}{T} \tag{22}$$

Derivation of the bias correction for the diagonal d'

The equation for the absolute value of the diagonal d' is presented in Equation 11. As in the case of signal modulations, for simplicity we proceed by estimating the bias for the squared diagonal d'. As above, we make the simplifying assumption that responses to each condition j are normally distributed with mean μ_j and variance σ_j^2, and that for each condition, μ_j is approximately equal to σ_j^2.

The numerator of the squared diagonal d' is given by the square of the difference between the mean match response and the mean distractor response. Since the response to each condition is assumed to be normally distributed, the difference between mean match and mean distractor is a linear combination of normal random variables and is also normally distributed. The numerator is equal to the square of this value, and it thus follows a non-central chi-squared distribution, whose mean is equal to the sum of the squared mean of the original normally distributed variable plus its variance:

$$Mean\left[\left(\mu_{Match}-\mu_{Distractor}\right)^2\right] = \left[Mean\left(\mu_{Match}-\mu_{Distractor}\right)\right]^2 + Variance\left(\mu_{Match}-\mu_{Distractor}\right) = \cdots$$

$$\cdots = \left(\sum_{i=1}^{4}\frac{1}{4}\cdot m_i - \sum_{i=1}^{12}\frac{1}{12}\cdot d_i\right)^2 + \frac{\sum_{i=1}^{4}\frac{1}{16}\cdot\sigma^2_{m_i,noise} + \sum_{i=1}^{12}\frac{1}{144}\cdot\sigma^2_{d_i,noise}}{T}$$

$$\cdots = \left(\sum_{i=1}^{4} \frac{1}{4} \cdot m_i - \sum_{i=1}^{12} \frac{1}{12} \cdot d_i \right)^2 + \frac{\sum_{i=1}^{4} \frac{1}{16} \cdot m_i + \sum_{i=1}^{12} \frac{1}{144} \cdot d_i}{T} \quad (23)$$

where m_i indicates the mean response to the i-th match and $\sigma^2_{m_i, noise}$ is its corresponding trial-by-trial variance, d_i indicates the mean response to the i-th distractor and $\sigma^2_{d_i, noise}$ is its corresponding trial-by-trial variance, and we used the assumption that the trial-by-trial variance for a given condition is equal to its corresponding mean response. The bias of the numerator of the squared d' is then equal to:

$$Bias\left[\left(\mu_{Match} - \mu_{Distractor} \right)^2 \right] = \frac{\sum_{i=1}^{4} \frac{1}{16} \cdot m_i + \sum_{i=1}^{12} \frac{1}{144} \cdot d_i}{T} \quad (24)$$

The denominator of the squared diagonal d' is equal to the pooled variance of the noise across matches and distractors. In the general case in which different conditions elicit different amounts of trial-by-trial variability, the denominator results in a linear combination of chi-squared variables, and its parameters can only be estimated in an approximated form (Satterthwaite 1946). However, one can note that the estimate of each individual trial-by-trial variance is unbiased, and therefore a linear combination of unbiased quantities is unbiased itself, thus no bias is introduced by the denominator. Consequently, the bias of the diagonal d' can be corrected by subtracting the bias of the squared numerator according to Equation 24, dividing it by the estimate of the pooled variance, and taking the square root (Equation 13).

Derivation of diagonal d' as a function of the orthonormal basis

Here we demonstrate that a neuron's diagonal d' can be deconstructed into a function of three "signal strengths" defined in terms of the orthonormal basis presented in Fig 1d. Diagonal d' is defined as the absolute value of the difference between the mean response to all target matches and the mean response to all distractors, divided by the pooled standard deviation of the noise (Equation 11).

The numerator of diagonal d' can thus be expressed as the absolute value of the dot product between the flattened response matrix $\boldsymbol{R}$ and a similarly formatted vector **c**, in which the target matches are scaled by $1/4$ and the distractors are scaled by $-1/12$:

$$\left|\mu_{Match} - \mu_{Distractor}\right| = \left|\sum_{i=1}^{4} \frac{1}{4} \cdot m_i - \sum_{i=1}^{12} \frac{1}{12} \cdot d_i\right| = \left|\boldsymbol{R} \cdot \boldsymbol{c}\right| \tag{25}$$

where m_i denotes the mean response to the i-th match and d_i denotes the mean response to the i-th distractor. The orthonormal basis function corresponding to the diagonal modulation $\boldsymbol{b}_{diag}$ is equal to $\boldsymbol{c}$ multiplied by $\sqrt{3}$ to impose unitary norm. As a result, the numerator of a neuron's diagonal d' can be rewritten as:

$$\left|\mu_{Match} - \mu_{Distractor}\right| = \left|\boldsymbol{R} \cdot \boldsymbol{c}\right| = \left|\boldsymbol{R} \cdot \frac{\boldsymbol{b}_{diag}}{\sqrt{3}}\right| = \sqrt{\frac{1}{3} \cdot w_{diag}^2} \tag{26}$$

The denominator of the diagonal d' is equal to the pooled standard deviation, i.e. the square root of the pooled variance (Equation 11). Our goal is to arrive at a

formulation of the pooled standard deviation as a function of the orthonormal basis weights.

We begin by expanding the terms for the variance of spike count responses to target matches σ^2_{Match} and to distractors $\sigma^2_{Distractor}$. If we indicate with m_{it} the response to the i-th match on the t-th trial, σ^2_{Match} can be rewritten as:

$$\sigma^2_{Match} = \frac{1}{80}\cdot\sum_{i=1}^{4}\sum_{t=1}^{20}\left(m_{it}-\mu_{Match}\right)^2 = \frac{1}{80}\cdot\sum_{i=1}^{4}\sum_{t=1}^{20}\left(m_{it}-m_i+m_i-\mu_{Match}\right)^2 = \cdots$$

$$= \frac{1}{4}\cdot\sum_{i=1}^{4}\left(m_i-\mu_{Match}\right)^2 + \frac{1}{4}\cdot\sum_{i=1}^{4}\frac{1}{20}\cdot\sum_{t=1}^{20}\left(m_{it}-m_i\right)^2 = \frac{1}{4}\cdot\sum_{i=1}^{4}\left(m_i-\mu_{Match}\right)^2 + \overline{\sigma}^2_{noise,Match} \tag{27}$$

where $\overline{\sigma}^2_{noise,Match}$ indicates the average trial-by-trial variability across the 4 matches, and m_i denotes the mean response to the i-th match. Similarly, if we indicate with d_{it} the response to the i-th distractor on the t-th trial, $\sigma^2_{Distractor}$ can be written as:

$$\sigma^2_{Distractor} = \frac{1}{240}\cdot\sum_{i=1}^{12}\sum_{t=1}^{20}\left(d_{it}-\mu_{Distractor}\right)^2 = \frac{1}{12}\cdot\sum_{i=1}^{12}\left(d_i-\mu_{Distractor}\right)^2 + \overline{\sigma}^2_{noise,Distractor} \tag{28}$$

where $\overline{\sigma}^2_{noise,Distractor}$ is the average trial-by-trial variability across the 12 distractors and d_i is the mean response to the i-th distractor. Now we substitute σ^2_{Match} and $\sigma^2_{Distractor}$ from equations 27 and 28 into equation 11, and express the pooled standard deviation as:

$$\sigma_{pooled} = \sqrt{\frac{1}{16}\cdot\left[\sum_{i=1}^{4}\left(m_i-\mu_{Match}\right)^2 + \sum_{i=1}^{12}\left(d_i-\mu_{Distractor}\right)^2\right] + \frac{4\cdot\overline{\sigma}^2_{noise,Match} + 12\cdot\overline{\sigma}^2_{noise,Distractor}}{16}} = \cdots$$

$$= \sqrt{\frac{1}{16} \cdot \left[\sum_{i=1}^{4} \left(m_i - \mu_{Match} \right)^2 + \sum_{i=1}^{12} \left(d_i - \mu_{Distractor} \right)^2 \right] + \bar{\sigma}^2_{noise}} = \sqrt{\sigma^2_{MD} + \bar{\sigma}^2_{noise}} \tag{29}$$

where $\bar{\sigma}^2_{noise}$ indicates the average trial-by-trial variability across all conditions (as defined in Equation 7), and σ^2_{MD} indicates the sum of the variance across matches and the variance across distractors:

$$\sigma^2_{MD} = \frac{1}{16} \cdot \left[\sum_{i=1}^{4} \left(m_i - \mu_{Match} \right)^2 + \sum_{i=1}^{12} \left(d_i - \mu_{Distractor} \right)^2 \right] \tag{30}$$

We now wish to express σ^2_{MD} as a function of the orthonormal basis components. Here we indicate the average response to the i-th condition as r_i and the grand mean spike count across all conditions as $\overline{SC}$, and we derive an expansion of the sum of the squared responses by substituting $\overline{SC} = \frac{1}{4} \cdot \mu_{Match} + \frac{3}{4} \cdot \mu_{Distractor}$:

$$\sum_{i=1}^{16} r_i^2 = \sum_{i=1}^{16} \left(r_i - \overline{SC} \right)^2 + 16 \cdot \overline{SC}^2 = \cdots$$

$$= \sum_{i=1}^{4} \left(m_i - \mu_{Match} + \mu_{Match} - \overline{SC} \right)^2 + \sum_{i=1}^{12} \left(d_i - \mu_{Distractor} + \mu_{Distractor} - \overline{SC} \right)^2 + 16 \cdot \overline{SC}^2 = \cdots$$

$$= \sum_{i=1}^{4} \left(m_i - \mu_{Match} \right)^2 + \sum_{i=1}^{12} \left(d_i - \mu_{Distractor} \right)^2 + 3 \cdot \left(\mu_{Match} - \mu_{Distractor} \right)^2 + 16 \cdot \overline{SC}^2 = \cdots$$

$$= 16 \cdot \sigma^2_{MD} + 3 \cdot \left(\mu_{Match} - \mu_{Distractor} \right)^2 + 16 \cdot \overline{SC}^2 \tag{31}$$

Equation 31 can be rearranged as:

$$\sigma_{MD}^2 = \frac{1}{16} \cdot \left[\sum_{i=1}^{16} r_i^2 - 16 \cdot \overline{SC}^2 - 3 \cdot \left(\mu_{Match} - \mu_{Distractor} \right)^2 \right] \tag{32}$$

The diagonal basis function $\boldsymbol{b}_{diag}$ and the grand mean basis function $\boldsymbol{b}_{mean}$ are defined such that their weights w_{diag} and w_{mean} take the following values:

$$w_{diag}^2 = \left(\boldsymbol{R} \cdot \boldsymbol{b}_{diag}^T \right)^2 = 3 \cdot \left(\mu_{Match} - \mu_{Distractor} \right)^2 \quad ; \quad w_{mean}^2 = \left(\boldsymbol{R} \cdot \boldsymbol{b}_{mean}^T \right)^2 = 16 \cdot \overline{SC}^2 \tag{33}$$

Because $\boldsymbol{b}_1 \ldots \boldsymbol{b}_{16}$ form an orthonormal basis,

$$\sum_{i=1}^{16} w_i^2 = \sum_{i=1}^{16} \left(\boldsymbol{R} \cdot \boldsymbol{b}_i^T \right)^2 = \sum_{i=1}^{16} r_i^2 \tag{34}$$

Substituting equations 33 and 34 into equation 32 allows us to derive:

$$\sigma_{MD}^2 = \frac{1}{16} \cdot \left[\sum_{i=1}^{16} w_i^2 - w_{diag}^2 - w_{mean}^2 \right] = \frac{1}{16} \cdot \sum_{\substack{i \neq diag, \\ i \neq mean}} w_i^2 \tag{35}$$

We now substitute equation 35 into equation 29:

$$\sigma_{pooled} = \sqrt{\frac{1}{16} \cdot \sum_{\substack{i \neq diag, \\ i \neq mean}} w_i^2 + \bar{\sigma}_{noise}^2} \tag{36}$$

Diagonal d' can thus be written as:

$$|d'| = \frac{\left| \mu_{Match} - \mu_{Distractor} \right|}{\sigma_{pooled}} = \sqrt{\frac{\frac{1}{3} \cdot w_{diag}^2}{\frac{1}{16} \cdot \sum_{\substack{i \neq diag, \\ i \neq mean}} w_i^2 + \bar{\sigma}_{noise}^2}} \tag{37}$$

In the raw formulation of the weights, rescaling a neuron's firing rate responses (e.g. multiplying a neuron's response matrix by two) results in a rescaling (i.e. a doubling) of all its deconstructed matrix weights, and consequently, modulations due to changes in the pattern of responses within the matrix and overall firing rates are entangled. To capture matrix structure in a manner that does not depend on the overall scaling of firing, we compute the "normalized weights" by dividing each weight by the grand mean spike count $\overline{SC}$. Dividing both numerator and denominator of Equation 37 by $\overline{SC}$ allows us to express diagonal d' as a function of the normalized weights:

$$|d'| = \sqrt{\frac{\frac{1}{3}\cdot w_{diag}^2}{\frac{1}{16}\cdot \sum_{\substack{i\neq diag,\\ i\neq mean}} w_i^2 + \bar{\sigma}_{noise}^2}} = \sqrt{\frac{\frac{1}{3}\cdot \left(\frac{w_{diag}}{\overline{SC}}\right)^2}{\frac{1}{16}\cdot \sum_{\substack{i\neq diag,\\ i\neq mean}} \left(\frac{w_i}{\overline{SC}}\right)^2 + \frac{1}{\overline{SC}}\cdot\left(\frac{\bar{\sigma}_{noise}^2}{\overline{SC}}\right)}} \tag{38}$$

Finally, we express diagonal d' as a function of three components:

$$|d'| = \sqrt{\frac{D}{ND + \frac{1}{\overline{SC}}}} \tag{39}$$

where:

$$D = \frac{1}{3}\cdot\left(\frac{w_{diag}}{\overline{SC}}\right)^2 \qquad ND = \frac{1}{16}\cdot \sum_{\substack{i\neq diag,\\ i\neq mean}} \left(\frac{w_i}{\overline{SC}}\right)^2 \qquad \frac{1}{\overline{SC}} = \frac{1}{\overline{SC}}\cdot\left(\frac{\bar{\sigma}_{noise}^2}{\overline{SC}}\right) \tag{40}$$

using the assumption that $\overline{SC}$ is equal to $\bar{\sigma}_{noise}^2$.

CHAPTER 5: Conclusions

In this dissertation we have examined the responses of populations of neurons recorded in IT and PRH as monkeys performed a DMS task. Our results showed that information about whether the currently viewed stimulus matches a sought target is more linearly separable in PRH than it is in IT, suggesting that PRH performs an "untangling" computation on the task-relevant information received from IT. We also showed that this linearly separable information arrives in PRH with a small delay after the initial arrival of nonlinearly separable information. Remarkably, a simple linear-nonlinear model could explain both these results. Finally, we presented a novel set of tools to directly analyze the responses of heterogeneous neural populations. In this chapter we discuss the implications of our results and some possible future directions.

The role of IT and PRH in target search

While previous studies of neural mechanisms during the DMS task pooled together the responses of neurons in both IT and PRH (Miller, Li et al. 1993, Miller and Desimone 1994, Miller, Erickson et al. 1996), in our work we have carefully distinguished the properties of the populations in these two brain areas. The differences we observed between the two populations were remarkable. Most importantly, we found evidence of an untangling computation performed by neurons in PRH. At the level of neural populations, this transformation was evidenced by a marked increase in the amount of linearly separable target match information from IT to PRH, whereas the total amount of

information was approximately equal in the two regions. At the level of single neurons, we observed a larger prevalence in PRH of neurons selective to specific target match conditions, which was best exemplified by the existence of a handful of “solution neurons” (i.e. neurons excited by all target match conditions, and suppressed by all distractor conditions or vice-versa) in PRH, while we did not find any such neurons in IT. Another difference between the two areas was a marked reduction in the amount of visual information in PRH, a fact that could be explained by our linear-nonlinear model.

Notably, significant differences between IT and PRH had been observed before in the context of visual search tasks. For example, in the context of a task requiring the memorization of specific pairs of objects (Miyashita 1988, Sakai and Miyashita 1991, Hirabayashi, Takeuchi et al. 2013), IT cells were found to be selective to individual objects, while PRH cells learned a more abstract representation that combined IT responses to express selectivity for specific pairs of objects. Moreover, significant differences between neural responses in IT and PRH have been observed in relation to different reward schemes during a DMS task (Liu and Richmond 2000). Finally, lesions to PRH are known to cause deficits in visual target search tasks only when the targets must be remembered over relatively long delays (above 1 minute), whereas lesions in IT impair performances even over extremely short delays (e.g. under 1 second) (Buffalo, Ramus et al. 1999, Buffalo, Ramus et al. 2000).

While our studies shed light on the role of IT and PRH during target search, a number of questions still remain open. First, while our results supported a scenario where all task-relevant information in PRH is inherited from IT, it is not clear whether IT is indeed the first stage of combination of visual and target signals, or whether they are first combined in some earlier area (e.g. V4) and later relayed to IT. Second, we

demonstrated that the representation of target match information was disrupted during error trials in IT and PRH. Although this is compatible with a role of these areas in the generation of behavior, we did not directly establish a causal link, thus leaving open other potential roles for the representation in PRH (e.g. to generate a reward prediction signal). Finally, it is not clear how our results might change in the context of more naturalistic conditions, in which objects are presented under a wide range of transformations (i.e. size, position and illumination changes), and where the number of possible objects is increased by several orders of magnitude (DiCarlo, Zoccolan et al. 2012).

Misaligned combinations of visual and target signals in IT and PRH

The neural responses we recorded in IT and PRH were extremely heterogeneous, and they featured a significant amount of "distractor detectors", as well as other "misaligned" combinations of visual and target signals. While in accord with previous reports (Haenny, Maunsell et al. 1988, Eskandar, Richmond et al. 1992), this result stands at odds with existing models of target search, which postulated that visual and target information are combined in the brain in a coordinated manner (Salinas 2004, Sugase-Miyamoto, Liu et al. 2008, Salinas and Bentley 2009, Engel and Wang 2011).

Why would the brain combine visual and target information in an apparently disorganized way? Here we propose two possible explanations. First, it might simply be too challenging to precisely route the top-down signals from PRH to individual IT cells, also considering the lack of topography in IT (Gawne and Richmond 1993). Instead, it might be more convenient to simply wire these connections at random in IT and to

reformat them at a later stage into an aligned format. Second, there might be a computational benefit in producing neurons with complex combinations of visual and target information. Even though such mixed selectivity is suboptimal to solve our particular DMS task, it has been postulated (Rigotti, Barak et al. 2013) that this coding approach might be advantageous to achieve the more general goal of offering a versatile representation to solve a large number of other unspecified tasks. For example, a particular non-match instance can provide useful context for the search by indicating where to look next to find the target. Future studies will be required to assess the actual usefulness of this representation during more complex tasks.

Interpretation of dynamic representations in the brain

Our results demonstrated the existence of a delay between the arrival of task-relevant information in PRH and the reformatting of this information into a more explicit representation. A similar evolution of neural responses has been reported before during visual search (Chelazzi, Miller et al. 1993, Chelazzi, Miller et al. 2001), as well as in the context of motion processing (Pack and Born 2001, Smith, Majaj et al. 2005) and object recognition (Brincat and Connor 2006). Previous studies have generally attributed this delay to the presence of recurrent circuits in the areas performing the computation. In this work we propose an alternative explanation: delays can naturally emerge from the action of instantaneous computations on a non-stationary input representation. Commonly reported non-stationarities, such as diversity in response latencies, are sufficient to evoke this phenomenon and can therefore play a significant role on downstream computation.

What is the source of the non-stationarities we measured in IT? Previous studies have reported complex dynamics in IT even in response to static visual images, and have demonstrated that these temporal fluctuations play a role in the encoding of visual information (Sugase, Yamane et al. 1999). At the same time, the neurons in PFC that are thought to maintain working memory information are also highly dynamic, and persistent activity is carried by different subpopulations at different times (Brody, Hernandez et al. 2003). A potential future direction of research to track the source of these non-stationarities might involve directly comparing their strength in IT and in the major inputs to this area (e.g. V4 and PFC).

Importantly, our model works under the assumption that its parameters were learned during a very specific time window after stimulus onset, namely one during which total information about the input was maximized. Although very little is known about the neural mechanisms underlying learning in PRH, we hypothesize that a potential implementation of our model could rely on the strong dopaminergic inputs to PRH (Richmond 2006), given their proposed role as a "time stamp" for the most salient moments during learning (Redgrave and Gurney 2006, Redgrave, Gurney et al. 2008).

Analysis of heterogeneous populations in high-level areas

A major element of our work was the development of a set of new quantitative analyses designed to probe the activity of heterogeneous neural populations. In particular, we designed a method to extract a set of intuitive signals from responses during complex tasks, and to map these signals onto measures of task performance. We predict that these tools will be particularly useful for the analysis of cognitive tasks, which

are composed by multiple “parameter axes” (e.g. object identity versus target identity), and for the study of neurons in high-level areas, given the extreme heterogeneity of their responses. A current limitation of our method consists of the lack of a mapping between the responses of single neurons and the population estimates of linearly separable information and total information. Future expansions will further elucidate the connections between single-neuron and population-based measurements of information.

Modeling untangling computations in the brain

Our work extended the “untangling” framework beyond object recognition (cox, dicarlo), thus demonstrating its applicability to cognition. Even though many tasks can be modeled within this framework, the mechanisms adopted by the brain to perform untangling computations are still unknown. In this work we made a first step in this direction by demonstrating that a simple linear-nonlinear model is sufficient to describe the computation between IT and PRH, and by providing an intuitive geometrical description of the underlying mechanism.

Because many tasks can be solved by untangling computations, a fascinating possibility is that the neural mechanisms underlying different untangling tasks share a similar “canonical” structure. From this perspective, it is notable that linear-nonlinear models analogous to the one we proposed have been successfully used to describe computations in multiple brain areas, including V1 (Rust, Schwartz et al. 2005), MT (Rust, Mante et al. 2006) and MST (Mineault, Khawaja et al. 2012). To provide a formalization of this intuition, we are currently developing a generalized version of our

linear-nonlinear model, such that it can be easily applied to recorded data in high-level areas, and whose parameters can be learned via biophysically plausible mechanisms. Future work will involve the application of such model to neural responses from different brain areas to test whether multiple neural computations can indeed be capture by a single mechanism.

BIBLIOGRAPHY

Adelson, E. H. and J. R. Bergen (1985). "Spatiotemporal energy models for the perception of motion." J Opt Soc Am A **2**(2): 284-299.

Adret, P., C. D. Meliza and D. Margoliash (2012). "Song tutoring in presinging zebra finch juveniles biases a small population of higher-order song-selective neurons toward the tutor song." J Neurophysiol **108**(7): 1977-1987.

Averbeck, B. B. and D. Lee (2006). "Effects of noise correlations on information encoding and decoding." J Neurophysiol **95**(6): 3633-3644.

Awh, E., K. M. Armstrong and T. Moore (2006). "Visual and oculomotor selection: links, causes and implications for spatial attention." Trends in cognitive sciences **10**(3): 124-130.

Barak, O. and M. Tsodyks (2014). "Working models of working memory." Current opinion in neurobiology **25**: 20-24.

Barak, O., M. Tsodyks and R. Romo (2010). "Neuronal population coding of parametric working memory." The Journal of Neuroscience **30**(28): 9424-9430.

Baxter, M. G. (2009). "Involvement of medial temporal lobe structures in memory and perception." Neuron **61**(5): 667-677.

Bennur, S. and J. I. Gold (2011). "Distinct representations of a perceptual decision and the associated oculomotor plan in the monkey lateral intraparietal area." J Neurosci **31**(3): 913-921.

Bichot, N. P., A. F. Rossi and R. Desimone (2005). "Parallel and serial neural mechanisms for visual search in macaque area V4." Science **308**(5721): 529-534.

Bishop, C. M. (2006). Pattern recognition and machine learning, springer New York.

Brendel, W., R. Romo and C. K. Machens (2011). Demixed principal component analysis. Advances in Neural Information Processing Systems 24.

Brincat, S. L. and C. E. Connor (2006). "Dynamic shape synthesis in posterior inferotemporal cortex." Neuron **49**(1): 17-24.

Brody, C. D., A. Hernandez, A. Zainos and R. Romo (2003). "Timing and neural encoding of somatosensory parametric working memory in macaque prefrontal cortex." Cerebral Cortex **13**(11): 1196-1207.

Buckley, M. J. and D. Gaffan (2006). "Perirhinal cortical contributions to object perception." Trends in cognitive sciences **10**(3): 100-107.

Buckley, M. J., F. A. Mansouri, H. Hoda, M. Mahboubi, P. G. Browning, S. C. Kwok, A. Phillips and K. Tanaka (2009). "Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions." Science **325**(5936): 52-58.

Buffalo, E. A., S. J. Ramus, R. E. Clark, E. Teng, L. R. Squire and S. M. Zola (1999). "Dissociation between the effects of damage to perirhinal cortex and area TE." Learning & Memory **6**(6): 572-599.

Buffalo, E. A., S. J. Ramus, L. R. Squire and S. M. Zola (2000). "Perception and recognition memory in monkeys following lesions of area TE and perirhinal cortex." Learn Mem **7**(6): 375-382.

Bussey, T. J. and L. M. Saksida (2005). "Object memory and perception in the medial temporal lobe: an alternative approach." Current opinion in neurobiology **15**(6): 730-737 %@ 0959-4388.

Carandini, M., J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant and N. C. Rust (2005). "Do we know what the early visual system does?" J Neurosci **25**(46): 10577-10597.

Chelazzi, L., J. Duncan, E. K. Miller and R. Desimone (1998). "Responses of neurons in inferior temporal cortex during memory-guided visual search." Journal of Neurophysiology **80**(6): 2918-2940.

Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (1993). "A neural basis for visual search in inferior temporal cortex." Nature **363**(6427): 345-347.

Chelazzi, L., E. K. Miller, J. Duncan and R. Desimone (2001). "Responses of neurons in macaque area V4 during memory-guided visual search." Cereb Cortex **11**(8): 761-772.

Chernick, M. R. (2007). Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition, Wiley.

Churchland, M. M., B. M. Yu, J. P. Cunningham, L. P. Sugrue, M. R. Cohen, G. S. Corrado, W. T. Newsome, A. M. Clark, P. Hosseini, B. B. Scott, D. C. Bradley, M. A. Smith, A. Kohn, J. A. Movshon, K. M. Armstrong, T. Moore, S. W. Chang, L. H. Snyder, S. G. Lisberger, N. J. Priebe, I. M. Finn, D. Ferster, S. I. Ryu, G. Santhanam, M. Sahani and K. V. Shenoy (2010). "Stimulus onset quenches neural variability: a widespread cortical phenomenon." Nat Neurosci **13**(3): 369-378.

Cohen, M. R. and J. H. Maunsell (2009). "Attention improves performance primarily by reducing interneuronal correlations." Nat Neurosci **12**(12): 1594-1600.

Curtis, C. E. and D. Lee (2010). "Beyond working memory: the role of persistent activity in decision making." Trends in cognitive sciences **14**(5): 216-222.

Dayan, P. and L. F. Abbott (2001). Theoretical Neuroscience, MIT Press.

Desimone, R. and J. Duncan (1995). "Neural Mechanisms of Selective Visual-Attention." Annual Review of Neuroscience **18**: 193-222.

DiCarlo, J. J. and D. D. Cox (2007). "Untangling invariant object recognition." Trends Cogn Sci **11**(8): 333-341.

DiCarlo, J. J., D. Zoccolan and N. C. Rust (2012). "How does the brain solve visual object recognition?" Neuron **73**(3): 415-434.

Edmonds, J. and E. L. Johnson (1970). Matching: a well-solved class of integer linear programs. Combinatorial structures and their applications: proceedings. R. K. Guy. Calgary, Gordon and Breach.

Efron, B. and R. J. Tibshirani (1994). An introduction to the boostrap. Boca Raton, CRC Press.

Engel, T. A. and X. J. Wang (2011). "Same or different? A neural circuit mechanism of similarity-based pattern match decision making." J Neurosci **31**(19): 6982-6996.

Eskandar, E. N., B. J. Richmond and L. M. Optican (1992). "Role of Inferior Temporal Neurons in Visual Memory .1. Temporal Encoding of Information About Visual Images, Recalled Images, and Behavioral Context." Journal of Neurophysiology **68**(4): 1277-1295.

Felleman, D. J. and D. C. Van Essen (1991). "Distributed hierarchical processing in the primate cerebral cortex." Cerebral cortex **1**(1): 1-47 1047-3211.

Froemke, R. C. and Y. Dan (2002). "Spike-timing-dependent synaptic modification induced by natural spike trains." Nature **416**(6879): 433-438.

Gaffan, D. (1974). "Recognition impaired and association intact in the memory of monkeys after transection of the fornix." Journal of comparative and physiological psychology **86**(6): 1100.

Gaffan, D. and E. A. Murray (1992). "Monkeys (< em> Macaca fascicularis) with rhinal cortex ablations succeed in object discrimination learning despite 24-hr intertrial intervals and fail at matching to sample despite double sample presentations." Behavioral neuroscience **106**(1): 30.

Gawne, T. J. and B. J. Richmond (1993). "How independent are the messages carried by adjacent inferior temporal cortical neurons?" The Journal of neuroscience **13**(7): 2758-2771.

Geisler, W. S. and D. G. Albrecht (1997). "Visual cortex neurons in monkeys and cats: detection, discrimination, and identification." Vis Neurosci **14**(5): 897-919.

Gibson, J. R. and J. H. R. Maunsell (1997). "Sensory modality specificity of neural activity related to memory in visual cortex." Journal of Neurophysiology **78**(3): 1263-1275.

Graf, A. B., A. Kohn, M. Jazayeri and J. A. Movshon (2011). "Decoding the activity of neuronal populations in macaque primary visual cortex." Nat Neurosci **14**(2): 239-245.

Grill-Spector, K., T. Kushnir, T. Hendler, S. Edelman, Y. Itzchak and R. Malach (1998). "A sequence of object-processing stages revealed by fMRI in the human occipital lobe." Human brain mapping **6**(4): 316-328.

Gu, Y., G. C. Deangelis and D. E. Angelaki (2012). "Causal links between dorsal medial superior temporal area neurons and multisensory heading perception." J Neurosci **32**(7): 2299-2313.

Haenny, P. E., J. H. R. Maunsell and P. H. Schiller (1988). "State Dependent Activity in Monkey Visual-Cortex .2. Retinal and Extraretinal Factors in V4." Experimental Brain Research **69**(2): 245-259.

Hirabayashi, T., D. Takeuchi, K. Tamura and Y. Miyashita (2013). "Microcircuits for hierarchical elaboration of object coding across primate temporal areas." Science **341**(6142): 191-195.

Holmes, E. J. and C. G. Gross (1984). "Stimulus equivalence after inferior temporal lesions in monkeys." Behavioral neuroscience **98**(5): 898.

Horel, J. A. (1996). "Perception, learning and identification studied with reversible suppression of cortical visual areas in monkeys." Behavioural brain research **76**(1): 199-214 %@ 0166-4328.

Horowitz, P., W. Hill and T. C. Hayes (1989). The art of electronics, Cambridge university press Cambridge.

Hung, C. P., G. Kreiman, T. Poggio and J. J. DiCarlo (2005). "Fast readout of object identity from macaque inferior temporal cortex." Science **310**(5749): 863-866.

Ito, M., H. Tamura, I. Fujita and K. Tanaka (1995). "Size and Position Invariance of Neuronal Responses in Monkey Inferoterotemporal Cortex."

Kelly, R. C., M. A. Smith, J. M. Samonds, A. Kohn, A. B. Bonds, J. A. Movshon and T. S. Lee (2007). "Comparison of recordings from microelectrode arrays and single electrodes in the visual cortex." J Neurosci **27**(2): 261-264.

Kobatake, E. and K. Tanaka (1994). "Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex." Journal of neurophysiology **71**: 856-856.

Lavenex, P., W. A. Suzuki and D. G. Amaral (2002). "Perirhinal and parahippocampal cortices of the macaque monkey: projections to the neocortex." J Comp Neurol **447**(4): 394-420.

Lehky, S. R. and K. Tanaka (2007). "Enhancement of object representations in primate perirhinal cortex during a visual working-memory task." Journal of neurophysiology **97**(2): 1298-1310.

Liebe, S., N. K. Logothetis and G. Rainer (2011). "Dissociable effects of natural image structure and color on LFP and spiking activity in the lateral prefrontal cortex and extrastriate visual area V4." J Neurosci **31**(28): 10215-10227.

Liu, Z. and B. J. Richmond (2000). "Response differences in monkey TE and perirhinal cortex: Stimulus association related to reward schedules." Journal of Neurophysiology **83**(3): 1677-1692.

Lueschow, A., E. K. Miller and R. Desimone (1994). "Inferior Temporal Mechanisms for Invariant Object Recognition." Cerebral Cortex **4**(5): 523-531.

Machens, C. K. (2010). "Demixing population activity in higher cortical areas." Front Comput Neurosci **4**: 126.

Machens, C. K., R. Romo and C. D. Brody (2005). "Flexible control of mutual inhibition: a neural model of two-interval discrimination." Science **307**(5712): 1121-1124.

Machens, C. K., R. Romo and C. D. Brody (2010). "Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex." J Neurosci **30**(1): 350-360.

Mansouri, F. A., M. J. Buckley and K. Tanaka (2007). "Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment." Science **318**(5852): 987-990.

Markov, N. T., M. M. Ercsey-Ravasz, A. R. R. Gomes, C. Lamy, L. Magrou, J. Vezoli, P. Misery, A. Falchier, R. Quilodran and M. A. Gariel (2012). "A weighted and directed interareal connectivity matrix for macaque cerebral cortex." Cerebral Cortex: 1047-3211.

Maunsell, J. H. R., G. Sclar, T. A. Nealey and D. D. Depriest (1991). "Extraretinal Representations in Area-V4 in the Macaque Monkey." Visual Neuroscience **7**(6): 561-573.

Maunsell, J. H. R. and S. Treue (2006). "Feature-based attention in visual cortex." Trends in neurosciences **29**(6): 317-322.

Merigan, W. H., T. A. Nealey and J. H. Maunsell (1993). "Visual effects of lesions of cortical area V2 in macaques." The Journal of neuroscience **13**(7): 3180-3191.

Meunier, M., J. Bachevalier, M. Mishkin and E. A. Murray (1993). "Effects on Visual Recognition of Combined and Separate Ablations of the Entorhinal and Perirhinal Cortex in Rhesus-Monkeys." Journal of Neuroscience **13**(12): 5418-5432.

Meyers, E. M., D. J. Freedman, G. Kreiman, E. K. Miller and T. Poggio (2008). "Dynamic population coding of category information in inferior temporal and prefrontal cortex." Journal of Neurophysiology **100**(3): 1407-1419.

Miller, E. K. and R. Desimone (1994). "Parallel Neuronal Mechanisms for Short-Term-Memory." Science **263**(5146): 520-522.

Miller, E. K., C. A. Erickson and R. Desimone (1996). "Neural mechanisms of visual working memory in prefrontal cortex of the macaque." Journal of Neuroscience **16**(16): 5154-5167.

Miller, E. K., L. Li and R. Desimone (1993). "Activity of neurons in anterior inferior temporal cortex during a short-term memory task." The Journal of neuroscience **13**(4): 1460-1478.

Mineault, P. J., F. A. Khawaja, D. A. Butts and C. C. Pack (2012). "Hierarchical processing of complex motion along the primate dorsal visual pathway." Proceedings of the National Academy of Sciences **109**(16): E972-E980.

Minsky, M. and S. Papert (1969). Perceptrons: An introduction to computational geometry. Cambridge, MA, MIT Press.

Mishkin, M. and J. Delacour (1975). "An analysis of short-term visual memory in the monkey." Journal of Experimental Psychology: Animal Behavior Processes **1**(4): 326.

Mishkin, M., E. S. Prockop and H. E. Rosvold (1962). "One-trial object-discrimination learning in monkeys with frontal lesions." Journal of Comparative and Physiological Psychology **55**(2): 178.

Miyashita, Y. (1988). "Neuronal correlate of visual associative long-term memory in the primate temporal cortex." Nature **335**(6193): 817-820.

Mongillo, G., O. Barak and M. Tsodyks (2008). "Synaptic theory of working memory." Science **319**(5869): 1543-1546.

Movshon, J. A., E. H. Adelson, M. S. Gizzi and W. T. Newsome (1985). The analysis of moving visual patterns. Pattern Recognition Mechanisms (Pontificiae Academiae Scientarium Scripta Varia. C. Chagas, R. Gattass and C. Gross. **54:** 117-151.

Murray, E. A. and T. J. Bussey (1999). "Perceptual–mnemonic functions of the perirhinal cortex." Trends in cognitive sciences **3**(4): 142-151.

Nakamura, K., K. Matsumoto, A. Mikami and K. Kubota (1994). "Visual response properties of single neurons in the temporal pole of behaving monkeys." Journal of Neurophysiology **71**: 1206-1206.

Naya, J., M. Yoshida, M. Takeda, R. Fujimichi and Y. Miyashita (2003). "Delay-period activities in two subdivisions of monkey inferotemporal cortex during pair association memory task. (vol 18, pg 2915, 2003)." European Journal of Neuroscience **18**(11): 3154-3154.

Naya, Y. and W. A. Suzuki (2011). "Integrating what and when across the primate medial temporal lobe." Science **333**(6043): 773-776.

Newsome, W. T., K. H. Britten and J. A. Movshon (1989). "Neuronal correlates of a perceptual decision." Nature **341**(6237): 52-54.

Pack, C. C. and R. T. Born (2001). "Temporal dynamics of a neural solution to the aperture problem in visual area MT of macaque brain." Nature **409**(6823): 1040-1042.

Pagan, M., U. L.S., W. M.P. and N. C. Rust (2013). "Signals in inferotemporal cortex and perirhinal cortex suggest an untangling of visual target information." Nature Neuroscience **16**: 1132-1139.

Pagan, M. and N. C. Rust (2014). "Quantifying the signals contained in heterogeneous neural responses and determining their relationships with task performance." Journal of Neurophysiology **112**(6): 1584-1598.

Panzeri, S., R. Senatore, M. A. Montemurro and R. S. Petersen (2007). "Correcting for the sampling bias problem in spike train information measures." J Neurophysiol **98**(3): 1064-1072.

Pillow, J. W., J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. J. Chichilnisky and E. P. Simoncelli (2008). "Spatio-temporal correlations and visual signalling in a complete neuronal population." Nature **454**(7207): 995-999.

Poor, H. V. (1994). An Introduction to Signal Detection and Estimation. New York, Springer.

Redgrave, P. and K. Gurney (2006). "The short-latency dopamine signal: a role in discovering novel actions?" Nat Rev Neurosci **7**(12): 967-975.

Redgrave, P., K. Gurney and J. Reynolds (2008). "What is reinforced by phasic dopamine signals?" Brain Res Rev **58**(2): 322-339.

Richmond, B. J. (2006). Dopamine-dependent associative learning of workload-predicting cutes in the temporal lobe of the monkey. Plasticity in the visual system: from genes to circuits. R. Pinaud, L. A. Tremere and P. De Weerd. New York, NY, Springer**:** 309-320.

Rigotti, M., O. Barak, M. R. Warden, X. J. Wang, N. D. Daw, E. K. Miller and S. Fusi (2013). "The importance of mixed selectivity in complex cognitive tasks." Nature **497**(7451): 585-590.

Rokni, D., V. Hemmelder, V. Kapoor and V. N. Murthy (2014). "An olfactory cocktail party: figure-ground segregation of odorants in rodents." Nature neuroscience: 1097-6256.

Romo, R., C. D. Brody, A. Hernandez and L. Lemus (1999). "Neuronal correlates of parametric working memory in the prefrontal cortex." Nature **399**(6735): 470-473.

Russ, B. E., A. L. Ackelson, A. E. Baker and Y. E. Cohen (2008). "Coding of auditory-stimulus identity in the auditory non-spatial processing stream." Journal of neurophysiology **99**(1): 87-95.

Rust, N. C. and J. J. DiCarlo (2010). "Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT." J Neurosci **30**(39): 12978-12995.

Rust, N. C., V. Mante, E. P. Simoncelli and J. A. Movshon (2006). "How MT cells analyze the motion of visual patterns." Nat Neurosci **9**(11): 1421-1431.

Rust, N. C., S. R. Schultz and J. A. Movshon (2002). "A reciprocal relationship between reliability and responsiveness in developing visual cortical neurons." J Neurosci **22**(24): 10519-10523.

Rust, N. C., O. Schwartz, J. A. Movshon and E. P. Simoncelli (2005). "Spatiotemporal elements of macaque v1 receptive fields." Neuron **46**(6): 945-956.

Sahgal, A., P. H. Galloway, I. G. McKeith, S. Lloyd, J. H. Cook, I. N. Ferrier and J. A. Edwardson (1992). "Matching-to-sample deficits in patients with senile dementias of the Alzheimer and Lewy body types." Archives of neurology **49**(10): 1043-1046.

Sakai, K. and Y. Miyashita (1991). "Neural organization for the long-term memory of paired associates." Nature **354**(6349): 152-155.

Salinas, E. (2004). "Fast remapping of sensory stimuli onto motor actions on the basis of contextual modulation." J Neurosci **24**(5): 1113-1118.

Salinas, E. and N. M. Bentley (2009). Gain modulation as a mechanism for switching reference frames, tasks, and targets. Coherent behavior in neuronal networks. K. Josic, J. Rubin, M. Matias and R. Romo. New York, Springer**:** 121-142.

Satterthwaite, F. E. (1946). "An approximate distribution of estimates of variance components." Biometrics Bulletin **2**: 110-114.

Schiller, P. H. (1995). "Effect of lesions in visual cortical area V4 on the recognition of transformed objects." Nature **376**(6538): 342-344.

Schmolesky, M. T., Y. Wang, D. P. Hanes, K. G. Thompson, S. Leutgeb, J. D. Schall and A. G. Leventhal (1998). "Signal timing across the macaque visual system." J Neurophysiol **79**(6): 3272-3278.

Schultz, W. (2007). "Behavioral dopamine signals." Trends Neurosci **30**(5): 203-210.

Schwartz, O., J. W. Pillow, N. C. Rust and E. P. Simoncelli (2006). "Spike-triggered neural characterization." J Vis **6**(4): 484-507.

Sharpee, T. O., C. A. Atencio and C. E. Schreiner (2011). "Hierarchical representations in the auditory cortex." Current opinion in neurobiology **21**(5): 761-767.

Shusterman, R., M. C. Smear, A. A. Koulakov and D. Rinberg (2011). "Precise olfactory responses tile the sniff cycle." Nature neuroscience **14**(8): 1039-1044.

Smith, M. A., N. J. Majaj and J. A. Movshon (2005). "Dynamics of motion signaling by neurons in macaque area MT." Nat Neurosci **8**(2): 220-228.

Stoerig, P. and A. Cowey (1997). "Blindsight in man and monkey." Brain **120**(3): 535-559.

Sugase, Y., S. Yamane, S. Ueno and K. Kawano (1999). "Global and fine information coded by single neurons in the temporal visual cortex." Nature **400**(6747): 869-873.

Sugase-Miyamoto, Y., Z. Liu, M. C. Wiener, L. M. Optican and B. J. Richmond (2008). "Short-term memory trace in rapidly adapting synapses of inferior temporal cortex." PLoS Comput Biol **4**(5): e1000073.

Suzuki, W. A. (1996). "The anatomy, physiology and functions of the perirhinal cortex." Current opinion in neurobiology **6**(2): 179-186.

Suzuki, W. A. and D. G. Amaral (1994). "Perirhinal and parahippocampal cortices of the macaque monkey: cortical afferents." J Comp Neurol **350**(4): 497-533.

Suzuki, W. A. and M. G. Baxter (2009). "Memory, perception, and the medial temporal lobe: a synthesis of opinions." Neuron **61**(5): 678-679.

Swaminathan, S. K. and D. J. Freedman (2012). "Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex." Nat Neurosci **15**(2): 315-320.

Tomita, H., M. Ohbayashi, K. Nakahara, I. Hasegawa and Y. Miyashita (1999). "Top-down signal from prefrontal cortex in executive control of memory retrieval." Nature **401**(6754): 699-703.

Treves, A. and S. Panzeri (1995). "The upward bias in measures of information derived from limited data samples." Neural Computation **7**: 399 - 407.

Tudusciuc, O. and A. Nieder (2007). "Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex." Proceedings of the National Academy of Sciences **104**(36): 14513-14518.

Wang, P. and D. Nikolic (2011). "An LCD Monitor with Sufficiently Precise Timing for Research in Vision." Front Hum Neurosci **5**: 85.

Yaginuma, S., T. Niihara and E. Iwai (1982). "Further evidence on elevated discrimination limens for reduced patterns in monkeys with inferotemporal lesions." Neuropsychologia **20**(1): 21-32.